

(12)

ARO Report 86-2

PROCEEDINGS OF THE THIRTY-FIRST CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING

AD-A169 473



Approved for public release; distribution unlimited.
The findings in this report are not to be construed
as an official Department of the Army position, un-
less so designated by other authorized documents.

Sponsored by
The Army Mathematics Steering Committee
on Behalf of

DTIC
ELECTE
JUL 2 1986
S A D

THE CHIEF OF RESEARCH, DEVELOPMENT AND ACQUISITION

86 7 1 070

DTIC FILE COPY

U. S. ARMY RESEARCH OFFICE

Report No. 86-2

June 1986

PROCEEDINGS OF THE THIRTY-FIRST CONFERENCE
ON THE DESIGN OF EXPERIMENTS

Sponsored by the Army Mathematics Steering Committee

HOST

Mathematics Research Center
University of Wisconsin
Madison, Wisconsin

23-25 October 1986

Approved for public release; distribution unlimited.
The findings in this report are not to be construed
as an official Department of the Army position, un-
less so designated by other authorized documents.

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

The Thirty-First Conference on the Design of Experiments in Army Research and Development and Testing was held 23-25 October 1985. The Army Mathematics Steering Committee (AMSC) is the sponsor of this series of meetings, and its subcommittee on Statistics and Probability organizes the scientific phase of each of them. Members of this subcommittee would like to thank Professor Bernard Harris for extending an invitation to hold this conference at the Mathematics Research Center, The University of Wisconsin, Madison, Wisconsin. His work, as chairperson for local arrangements, was a big factor in the success of this meeting.

This year eighteen contributed papers were given in the clinical and technical sessions. Most of these were presented by Army scientists. The titles of the sessions give some indication of the statistical areas treated: (1) Final Series and Multivariate Analysis, (2) Consistence Analysis, (3) Experimental Design, (4) Statistical Modeling, (5) Data Analysis, (6) Reliability and Quality Control. For the invited speaker phase of the conference, the Program Committee was pleased to obtain the services of the following nationally known scientists to talk on topics of current interest to Army personnel:

Speaker and Affiliation

Titles of Address

Professor Jerome Sacks
University of Illinois at
Urbana-Champaign

Keynote Address

Professor Marion R. Reynolds, Jr.
Virginia Polytechnic Institute
and State University

Approaches to Statistical
Validation of Simulation Models

Dr. Daryl Pregibon
Bell Laboratories

An Expert System for Data
Analysis

Dr. Howard Wainer
Educational Testing Services

How to Display Data Badly

Professor Gouri K. Bhattacharyya

Accelerated Life Tests

Since the Army analytic community is becoming ever more involved in the use of expert opinion and the related approaches to the analysis of new systems performance measures, it seemed an ideal time to have a special session to provide the audience with new insight into this important area. The AMSC is indebted to Professor Nazer D. Singpurwalla of George Washington University for organizing and chairing this feature session entitled, "Using Expert Opinions and Expert Systems in Reliability and Maintainability". We note below the titles of the addresses given by the four speakers in this informative session.

HUMAN FACTORS AFFECTING SUBJECTIVE JUDGMENTS

Mary A. Meyer, Energy Technology Group, Los Alamos National Laboratories

SOURCES AND EFFECTS OF CORRELATION OF EXPERT OPINIONS

Jane M. Booker, Statistics Group, Los Alamos National Laboratories

USE OF EXPERT OPINION IN RELIABILITY ASSESSMENT OF THE M-1 ABRAMS TANK

Bobby Bennett, U.S. Army Material Systems Analysis Agency

A MATHEMATICAL THEORY OF TESTABILITY

Alan Currit, Systems Product Division, IBM, Rochester

Professor Emanuel Parzen, Department of Statistics at Texas A&M University was selected by the AMSC to receive the Fifth Wilks Award for Contributions to Statistical Methodologies in Army Research Development and Testing. He richly deserves this honor for his many significant contributions to time series modeling and analysis, stochastic processes, statistical theory (including his seminal paper on density estimation), and his recent work on the foundations and generalized methodologies in data analysis. His latest work will undoubtedly have a very pronounced effect on the theory and practice of statistics in the years to come.

The AMSC has requested that the proceedings of the 1985 conference be distributed Army-wide so that the information contained therein can assist scientists with some of their statistical problems. Finally, committee members would like to thank the Program Committee for all the work it did in putting together this scientific meeting.

PROGRAM COMMITTEE

Carl Bates
Robert Burge
Bernard Harris
Robert Launer

William McIntosh
J. Richard Moore
Douglas Tang
Malcolm Taylor

Jerry Thomas

Accession For	
NTIS GRA&I	<input checked="checked" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
AI	

→ Partial

TABLE OF CONTENTS*

<u>Title</u>	<u>Page</u>
Foreword	111
Table of Contents	v
Program	vi1
→ APPROACHES TO STATISTICAL VALIDATION OF SIMULATION MODELS;	
Marion R. Reynolds, Jr.	1
→ DISTRIBUTION UNDER DEPENDENCE OF NONPARAMETRIC TWO-SAMPLE TESTS;	
Emanuel Parzen	19
→ STATISTICAL MODELS AND METHODS FOR CLUSTER ANALYSIS AND SEGMENTATION;	
Stanley L. Sclove	29
→ A COMPARISON OF METHODS FOR FACTOR ANALYSIS OF VISIBILITY;	
Oskar M. Essenwanger	39
→ SMALL COMPOSITE DESIGN;	
Norman R. Draper	61
→ CONSIDERATIONS IN SMALL SAMPLE QUANTAL RESPONSE TESTING;	
Barry A. Bodd and Henry B. Tingey	65
→ HUMAN FACTORS AFFECTING SUBJECTIVE JUDGMENTS;	
Mary A. Meyer	85
→ USE OF EXPERT OPINION IN THE RELIABILITY ASSESSMENT OF THE M-1 ABRAMS TANK;	
Bobby G. Bennett	99

→ VI
P

* This Table of Contents contains only the papers that are published in this technical manual. For a list of all papers presented at the Thirtieth Conference on the Design of Experiments, see the Program of this meeting.

Fr. P. V

Title

Page

→ APPLICATION OF HYPOTHESIS TESTING TO PERFORMANCE APPRAISAL ;

Richard M. Duncan and Paul H. Thrasher 107

→ MODELS FOR CONTINGENCY TABLE DATA ;

R. A. Kolb 145

→ ON A CLASS OF PROBABILITY DENSITY FUNCTIONS ;

H. P. Dudel and S. H. Lehnigk 165

→ PLOTTING MATHEMATICAL FUNCTIONS ON A STANDARD LINE PRINTER ;

Donald W. Rankin 189

→ STATISTICAL COMPARISON OF THE ABILITY OF CAMOUFLAGE COLORS TO BLEND WITH TERRAIN BACKGROUND UNDER HIGH AND LOW SUN ANGLES ;

George Anitole, Ronald I. Johnson and Christopher J. Neubert . . . 201

→ WEIBULL TAIL MODELING FOR ESTIMATING CONFIDENCE ON QUANTILES FROM CENSORED SAMPLES ;

Mark Vangel 213

→ THE LINDSTROM-MADDEN METHOD FOR SERIES SYSTEMS WITH REPEATED COMPONENTS ;

Andrew P. Soms 229

→ HOW TO DISPLAY DATA BADLY ; and

Howard Wainer 241

→ ACCELERATED LIFE TEST: AN OVERVIEW AND SOME RECENT ADVANCES

Gouri K. Bhattacharyya 253

Roster of Attendees 275

AGENDA

**THIRTY-FIRST CONFERENCE ON THE DESIGN OF EXPERIMENTS
IN ARMY RESEARCH, DEVELOPMENT AND TESTING**

23-25 October 1985

Host: The Mathematics Research Center

**Location: The Wisconsin Center Wisconsin Memorial Union
702 Langdon Street Langdon & Park Streets
Madison, Wisconsin (parallel sessions)**

******* Wednesday, 23 October *******

0815-0915 REGISTRATION - First Floor, The Wisconsin Center

0915-0930 CALLING OF THE CONFERENCE TO ORDER

Lake Shore Room, The Wisconsin Center

Prof. Bernard Harris, The Mathematics Research Center

WELCOMING REMARKS

0930-1200 GENERAL SESSION I - Lake Shore Room, The Wisconsin Center

Chairman: Prof. Bernard Harris

0930-1030 KEYNOTE ADDRESS

Jerome Sacks, University of Illinois at Urbana-Champaign

1030-1100 BREAK

1100-1200 APPROACHES TO STATISTICAL VALIDATION OF SIMULATION MODELS

**Marion R. Reynolds, Jr., Virginia Polytechnic Institute
and State University**

1200-1330 LUNCH

1330-1500 TECHNICAL SESSION I; TIME SERIES AND MULTIVARIATE ANALYSIS

Lake Shore Room, The Wisconsin Center

Chairman: William D. Baker, Ballistic Research Laboratory

**TESTS OF EQUALITY OF DISTRIBUTIONS FOR DEPENDENT SAMPLES
AND STATIONARY TIME SERIES**

Emanuel Parzen, The Texas A&M University

**ON SEGMENTATION OF SIGNALS, TIME SERIES, AND IMAGES:
IMPROVED ESTIMATION AND SEQUENTIAL PROCESSING**

Stanley Solove, The University of Illinois at Chicago

A COMPARISON OF METHODS FOR FACTOR ANALYSIS OF VISIBILITY

Oskar M. Essenwanger, US Army Missile Command

1300-1330 BREAK

1530-1630 CLINICAL SESSION

Old Madison Room, Wisconsin Memorial Union, 3rd Floor

Chairman: Oskar M. Essenwanger, US Army Missile Command

Panelists:

Prof. Bernard Harris, The Mathematics Research Center

Prof. Richard Johnson, The University of Wisconsin - Madison

Prof. Stanley Solove, The University of Illinois at Chicago

CONSISTENCY ANALYSIS OF AUTOMATIC TARGET RECOGNIZER PERFORMANCE

Clarence P. Walters, Night Vision and Electro-Optics Lab

1530-1700 TECHNICAL SESSION II; EXPERIMENTAL DESIGN

Inn Wisconsin Room, Wisconsin Memorial Union, 2nd floor

Chairman: Carl Bates, US Army Concepts Analysis Agency

SMALL COMPOSITE DESIGNS

Norman R. Draper, The University of Wisconsin

**HIGH FREQUENCY RADIO GROUND COMMUNICATIONS: DESIGNING TESTS
FOR 1980's APPLICATIONS OF 1940's TECHNOLOGY**

Clarence H. Annett, TRADOC Independent Evaluation Directorate

CONSIDERATIONS IN SMALL SAMPLE QUANTAL RESPONSE TESTING

Barry A. Bodd, Ballistic Research Laboratory

Henry B. Tingey, The University of Delaware

1830-1930 CASH BAR

1930-2130 BANQUET AND PRESENTATION OF WILKS AWARD

The Howard Johnson's Executive Hotel
525 West Johnson Street
Madison, WI
(The Conference Hotel)

* * * * * Thursday, 24 October * * * * *

0830-1030 SPECIAL SESSION ; USING EXPERT OPINIONS AND EXPERT SYSTEMS
IN RELIABILITY AND MAINTAINABILITY

Lake Shore Room, The Wisconsin Center

Chairman and Coordinator: Moxer D. Singpurwalla, The George
Washington University

HUMAN FACTORS AFFECTING SUBJECTIVE JUDGEMENTS

Mary A. Meyer, Energy Technology Group, Los Alamos National Laboratories

SOURCES AND EFFECTS OF CORRELATION OF EXPERT OPINIONS

Jane M. Booker, Statistics Group, Los Alamos National Laboratories

USE OF EXPERT OPINION IN RELIABILITY ASSESSMENT OF THE M-1 ABRAMS TANK

Bobby Bennett, U.S. Army Material Systems Analysis Agency

A MATHEMATICAL THEORY OF TESTABILITY

Alan Currit, Systems Product Division, IBM, Rochester

1030-1100 BREAK

1100-1200 GENERAL SESSION II - Lake Shore Room, The Wisconsin Center

Chairman: Malcolm Taylor, Ballistic Research Laboratory

TITLE: TO BE ANNOUNCED (An Expert System for Data Analysis)

Daryl Pregibon, Bell Laboratories

1200-1330 LUNCH

1330-1500 TECHNICAL SESSION III; STATISTICAL MODELING

Old Madison Room, Wisconsin Memorial Union 3rd floor

Chairman: Richard L. Umholtz, Ballistic Research Laboratory

APPLICATION OF HYPOTHESIS TESTING TO PERFORMANCE APPRAISAL

Richard H. Duncan, Technical Director, and Chief Scientist
White Sands Missile Range

Paul H. Thrasher, White Sands Missile Range

MODELS FOR CONTINGENCY TABLE ANALYSIS

Rickey A. Kolb, United States Military Academy

A CLASS OF PROBABILITY DENSITY FUNCTIONS

Siegfried H. Lehnigk, US Army Missile Range

PLOTTING MATHEMATICAL FUNCTIONS ON A STANDARD LINE PRINTER

Donald W. Rankin, LtCol, USAF, Ret, El Paso

1330-1500 TECHNICAL SESSION IV; DATA ANALYSIS

Beefeaters Room, Wisconsin Memorial Union, 3rd floor

Chairman: James C. Ford, Ballistic Research Laboratory

QUANTITATIVE ASSESSMENT OF THE INTERACTION AND ACTIVITY OF COMBINATIONS
OF ANTIPARASITIC DRUGS IN CONTINUOUS IN VITRO CULTURE OF
PLASMODIUM FALCIPARUM

Robert E. Miller, Walter Reed Army Institute of Research

STATISTICAL ANALYSIS OF PAVEMENT EVALUATION DATA

Starr D. Kohn, Waterways Experiment Station
Walter R. Barker, Waterways Experiment Station

STATISTICAL COMPARISON OF THE ABILITY OF CAMOUFLAGE COLORS TO BLEND WITH
TERRAIN BACKGROUND UNDER HIGH AND LOW SUN ANGLES

George Anitole, US Army Belvoir Research & Development Center
Ronald L. Johnson, US Army Belvoir Research & Development Center

1500-1530 BREAK

1530-1700 TECHNICAL SESSION V; RELIABILITY AND QUALITY CONTROL

Old Madison Room, Wisconsin Memorial Union, 3rd floor

Chairman: Donald Neal, Army Materials and Mechanics Research Center

THE LINDSTROM-MADDEN METHOD FOR SERIES SYSTEMS WITH REPEATED COMPONENTS

Andrew P. Soms, The University of Wisconsin-Milwaukee

CONVERTING INDIVIDUAL SAMPLING PLANS TO A COMPARABLE GROUP PLAN

Paul A. Roediger, US Army Armament, Munitions and Chemical Command

John A. Mardo, US Army Armament, Munitions and Chemical Command

**WEIBULL EXTREME QUANTILE MODELING FOR ESTIMATING CONFIDENCE ON
RELIABILITY FROM CENSORED SAMPLES**

Mark Vangel, Army Materials and Mechanics Research Center

AN ALGORITHM FOR DIAGNOSIS OF SYSTEM FAILURE

Robert L. Launer, US Army Research Office

*** * * * * Friday, 25 October * * * * ***

0900-1200 GENERAL SESSION III - Lake Shore Room, The Wisconsin Center

**Chairman: Douglas B. Tang, Walter Reed Army Institute of Research
Chairman of the AMSC Subcommittee on Probability and Statistics**

**0900-0930 OPEN MEETING OF THE STATISTICS AND PROBABILITY SUBCOMMITTEE OF THE
ARMY MATHEMATICS STEERING COMMITTEE**

0930-1030 HOW TO DISPLAY DATA BADLY

Howard Wainer, Educational Testing Service

1030-1100 BREAK

1100-1200 ACCELERATED LIFE TESTS: AN OVERVIEW AND SOME RECENT ADVANCES

Gouri K. Bhattacharyya, The University of Wisconsin-Madison

ADJOURN

APPROACHES TO STATISTICAL VALIDATION OF SIMULATION MODELS

Marion R. Reynolds, Jr.
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061

ABSTRACT

The process of validating a stochastic simulation model usually involves the comparison of data generated by the model with corresponding data from the real system. One method of making this comparison is to test the hypothesis that the distribution of model output is the same as the distribution of the corresponding variable in the real system. Since no model is a perfect reflection of the real system, a more realistic formulation is to test the hypothesis that the model is close enough for the purposes of the model user. An alternate approach to validation considers the error that results when the model is used to predict the behavior of the real system. In order to help the model user evaluate the predictive ability of the model, confidence intervals for expected error or prediction intervals for actual error can be constructed.

1. SIMULATION MODELS

Stochastic simulation models are now widely used in many fields to model complex systems when other types of models can not be used. In many cases the system being modeled will include many simpler processes interacting in a dynamic setting so that it is not possible to carry through a direct mathematical analysis. The nature of a simulation model usually means that the basic assumptions and structure of the model are not readily apparent to the model user so that model validation is particularly important for these models.

Models can be constructed for several purposes, for example to gain basic understanding of the system being modeled, to compare different management strategies with the idea of selecting a good strategy, or to predict the behavior of the system being modeled. In each of these cases some inference obtained using the model will be applied to the real system. In most situations the ability of the model to predict system behavior will be critical to the effectiveness of the model. The main purpose of the model will usually determine the predictive ability required of the model and this in turn will influence the approach to validation that is required.

2. VALIDATION

Before a simulation model can be used with confidence, the model user needs to know whether the model is a reasonable representation of the real system so that inferences or predictions obtained from the model are useful for the real system. It is the need for this type of information that leads to issues of validation and assessment of the model.

In discussing model validation it is usually not helpful to think in absolute terms of a model being either valid or invalid, but rather in terms

of degree of validity or, better yet, in terms of degree of usefulness. The usefulness of a model will depend on the purpose of the model and on the conditions under which it is used. For example, a model may be useful for determining the relative performance of two management strategies but not very useful for providing accurate and detailed predictions of future system behavior. A model which is useful for providing predictions for 5 years in the future may not provide useful predictions for 15 years in the future.

A useful way to think about the nature of validation has been given by Van Horn (1971). He defined validation as "the process of building an acceptable level of confidence that an inference about a simulated process is a correct or valid inference for the actual process". An important point here is that validation is a process and not a one time exercise. Ideally, the validation process should be carried out during the model building process (Sargent (1979)) as well as after the model is essentially complete. Another important point in Van Horn's definition is that validation is a process of building confidence in the model and not necessarily the process of "proving" that the model is valid.

It may be helpful to make a distinction between validation and what Fishman and Kiviat (1968) have called verification. Verification is the process of determining whether the simulation model behaves as the model builders intended. For example, "debugging" the computer program is an important part of the verification process. The validation process extends beyond the verification process since a model which behaves exactly as the model builders intended still may not be useful for drawing inferences about the real system.

3. APPROACHES TO VALIDATION

Some of the discussion of validation in the simulation literature has focused on philosophical issues. Discussion of some of the issues involved are given in McKenney (1967), Naylor and Finger (1967), Schrank and Holt (1967), and Shannon (1975). Balci and Sargent (1984) give an up-to-date bibliography of papers dealing with various aspects of model validation.

One direct approach to validation involves examining the model for "face validity", that is, determining whether the assumptions and structure of the model seem reasonable to people who are knowledgeable about the real system (see, for example, Law (1982)). This examination of assumptions should, of course, be carried out during the modeling process as the modeler develops a conceptual model in collaboration with people who are familiar with the system. After the model has been constructed other "independent" experts can be used to evaluate the model.

In addition to examining assumptions for conformance to existing knowledge and theory, empirical testing of these assumption can be carried out (Naylor and Finger (1967)). In this context the use of sensitivity analysis may help to identify which assumptions are most critical so that attention can be focused on these critical assumptions (Van Horn (1972)). In addition to a sensitivity analysis conducted in the likely range of model parameters, an evaluation of model performance can be done at the extremes of the parameter values (Sargent (1983)).

One of the most important tests to which a model can be subjected in the validation process is the comparison of data obtained from the real system with corresponding data generated from the model. If there is close agreement, in some sense, between these two data sets then this will increase confidence in the model. Some authors argue that the ability of the model to

predict the behavior of the real system is the most important test of a model.

Confidence in the model will be higher when the data used in the validation of the model is independent of the data used in constructing the model. If it is not possible to obtain separate data for validation then one approach is to split the existing data into two sets. One set can be used for constructing the model and the other set can be used for validating the model. In many cases the data used in constructing and validating a model will be historical data that has been collected on the existing system or a similar system. Ideally the model should be tested by its ability to predict the behavior of the system in the future. This may not be immediately possible either because the real system may not yet exist or because there is not enough time to wait for future observations on the real system. This paper will concentrate on the case where validation data is available since this is the case where statistical approaches can be used in comparing the model and the real system.

4. EXAMPLE

When discussing various statistical techniques that are useful in validation it may be helpful to think in terms of a specific type of simulation model as an example. Consider the model PTAEDA developed by Daniels and Burkhardt (1975) for simulating the growth of trees in forest stands. This type of model is designed to model stand growth over time so that various management strategies or the effects of various natural phenomena can be evaluated. The volume of wood in a stand at some future time is one of the main system variables of interest, but other variables such as the number of trees in various diameter classes may also be of

interest. In this model individual trees within the stand are assigned initial coordinate locations and sizes at an age corresponding to the onset of competition. Then annual diameter and height growth of each tree is simulated as a function of tree size, site quality, age, and an index reflecting competition from neighboring trees. Tree growth is adjusted by a random component representing genetic and/or microsite variability. Each year each tree survives with a certain probability and this survival probability is a function of tree size and competition. The wood volumes for individual trees at the end of the simulation period are obtained by substituting diameter and height values into tree volume equations. Estimates of wood yield per unit area are obtained by summing the individual tree volumes and multiplying by an appropriate expansion factor.

5. NOTATION

Suppose that the simulation model is constructed in such a way that p input variables represented by $\underline{X} = (X_1, X_2, \dots, X_p)$ are used to generate an output variable represented by Z . The input variables are usually selected to correspond to the most important observable input variables in the real system. The output variable Z in the model corresponds to some variable Y that is of interest in the real system. For example, for a forest stand simulator designed to predict stand volume at a future time, \underline{X} might represent input variables such as site quality, stand age at the future time, and some measure of current density. Z would correspond to simulated stand volume from the model and Y would correspond to the actual stand volume at the future time. In most applications it will be reasonable to treat both Y and Z as random variables whose distributions depend on the levels of \underline{X} . Y is a random variable because the value of Y can not be determined by determining the values of a finite number of input variables and Z is of

course a random variable because the model contains stochastic elements. Since the distributions of Y and Z depend on \underline{X} it will be convenient to work with $F(y|\underline{x})$ and $G(z|\underline{x})$, the conditional distribution functions of Y and Z , respectively.

Model users will usually be interested in using a model to make two general types of inferences about the real system being modeled. The first type of inference is concerned with a parameter or characteristic associated with the distribution of the variable Y from the real system. The parameter that is usually of most interest is the conditional mean $E(Y|\underline{x})$; other parameters that might be of interest are $P(Y \leq y|\underline{x})$, the probability that the system output is below a specified value y , and the variance $\text{Var}(Y|\underline{x})$. All of these parameters are functions of the input variables \underline{X} . For example a model user might be interested in estimating the average volume for stands of a particular type where the type of stand is determined by specifying the input variables age, site quality, and density. Alternately, the user might want to estimate the probability that a stand of a particular type has a volume below an economically determined lower threshold.

The second type of inference is concerned with predicting an actual value of Y that is to be observed when \underline{X} is at some specified value. For example the model user might be interested in a particular stand and want to predict the volume on this stand (as opposed to the average volume on all stands of this type). The usefulness of the model for making either type of inference depends on how close the conditional distribution of Z , given $\underline{X} = \underline{x}$, is to the conditional distribution of Y , given $\underline{X} = \underline{x}$. The best that could be hoped for is that these two conditional distributions are equal. Even then, in any trial of the model, the simulated value of Z will not necessarily be close to the corresponding observed value of Y since both Z

and Y are random variables.

Suppose that observations from the real system are available for n different sets of conditions, and for the i th set of conditions m_i observations from the real system are available. Let

Y_{ij} = j th observation from the real system under the i th set of conditions

and

$$\underline{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{im_i}) .$$

For example, data on total wood volume may be available for n different types of plots. In this example each plot may be distinct so that $m_i = 1$ for all i . Also let

$$\underline{X}_i = (X_{i1}, \dots, X_{ip})$$

= input variables for the i th set of conditions.

Corresponding to the i th set of conditions represented by $\underline{X}_i = \underline{x}_i$, the simulation model can be run m'_i times to generate m'_i independent simulated values which can be represented by

$$\underline{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{im'_i}) .$$

In some cases it may be useful to use the components of \underline{Y}_i and \underline{Z}_i individually, but in other cases the averages may be used. Then

$$\bar{Y}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Y_{ij}$$

is an estimator of $E(Y|\underline{x}_i)$, the mean of the system at the i th set of

conditions, and

$$\bar{Z}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Z_{ij}$$

is an estimator of $E(Z|\underline{x}_i)$, the mean of the model at the i th set of conditions. The bias or expected error in the model at $\underline{X}_i = \underline{x}_i$ is $E(Y-Z|\underline{x}_i)$ and an unbiased estimator of this bias is

$$D_i = D(\underline{x}_i) = \bar{Y}_i - \bar{Z}_i .$$

It may also be useful to think of \bar{Z}_i as a predictor of \bar{Y}_i before \bar{Y}_i is observed and in this case D_i is the prediction error.

6. HYPOTHESIS TESTING

In developing a model based on a finite number of input variables \underline{X} , the best model that could be achieved would have the conditional distribution of Z given $\underline{X} = \underline{x}$ equal to the conditional distribution of Y given $\underline{X} = \underline{x}$. Thus a natural way to formulate the validation problem is as the problem of testing the null hypothesis that Z and Y have the same conditional distributions. Let A be a set representing the range of input variables for which it is desirable to validate the model. Then the problem can be stated formally as one of testing

$$H_0: F(\cdot|\underline{x}) = G(\cdot|\underline{x}) \text{ for all } \underline{x} \in A .$$

The alternative is that F and G are not equal for at least one $\underline{x} \in A$. Ideally the set of validation data should be representative of A in some way, for example, a random sample from A . In practice it may not be feasible to

take a random sample and thus whatever data is available may have to be used. For purposes of building confidence in the model, data that represents the extremes of A might actually be better than a random sample. If the validation data does not adequately cover A then of course the conclusions about model validity that can be drawn from the data would be restricted to the subset of A represented by the data.

A reasonable interpretation of the hypothesis testing formulation of the validation problem is that the test is being carried out to determine whether there is any indication that the model does not represent the real system. If the null hypothesis is not rejected then this is interpreted to mean that there is no strong evidence of model inadequacy. It does not of course mean that the model is a perfect reflection of the real system or that the model can not be improved upon since the power of the test used may not be high. On the other hand a decision to reject the null hypothesis does not necessarily mean that the model is not useful. Rejection in this case would be taken as an indication that there is room for improvement and that the data should be examined for indications of areas for model improvement.

In some cases the requirement that F and G be equal may be too strict and a test for equal conditional means may be sufficient. In this case the null hypothesis would be

$$H_0: E(Y|\underline{x}) = E(Z|\underline{x}) \text{ for all } \underline{x} \in A.$$

If m and m' are small there may not be enough information at the set of conditions represented by $\underline{x} = \underline{x}_1$ to provide a test of either H_0 or H_0' with reasonable power. In this case it would be reasonable to apply a test at each set of conditions and then use some method for combining independent

tests. One well known method of combining independent test was developed by Fisher (1938). Let T_i be the test that is applied at the i th set of conditions and let α_i represent the observed significance level of the test, i.e. α_i is the probability of a value of T_i that is as extreme or more extreme than the observed value of T_i . If the distribution of T_i is continuous then the distribution of α_i is uniform on $(0,1)$ when the null hypothesis is true. From this it can be shown that $-2 \sum_{i=1}^n \log \alpha_i$ has a chi-square distribution with $2n$ degrees of freedom when the null hypothesis is true. When the α_i are small, $-2 \sum_{i=1}^n \log \alpha_i$ will be large and Fisher's test rejects the null hypothesis when $-2 \sum_{i=1}^n \log \alpha_i$ exceeds an appropriate critical value from the chi-square table. For other methods of combining independent tests see, for example, Osterhoff (1969). Alternately, a procedure such as the analysis of variance could be used to combine information if the usual assumptions such as equality of variances at the different conditions are reasonable.

7. CHOICE OF A TEST

For testing H_0 a test such as the two-sample Kolmogorov-Smirnov test for the equality of two distribution functions could be used. This test could be applied to \underline{Y}_i and \underline{Z}_i at each set of conditions and then information from all tests could be combined together. This type of test has the disadvantage that it is designed for the very general alternative $F(\cdot|\underline{x}) \neq G(\cdot|\underline{x})$ for some $\underline{x} \in A$ and thus may not have high power for specific alternatives that may be of primary interest.

For testing H_0' various parametric and nonparametric tests could be used. If normality and constant variance can be assumed then the analysis of variance is a reasonable choice where there are two treatments (real and simulation) and n blocks corresponding to the n sets of conditions. If con-

stant variance can not be assumed then individual two-sample t-statistics can be computed at each point and then combined into an overall test. If H_0' is rejected then the individual t-statistics would be useful in indicating places where the model does not work well.

If normality can not be assumed then two-sample nonparametric tests such as the Wilcoxon rank sum test can be used at each point and combined into an overall test. In many applications data on the real system may be scarce and there may be only one real observation Y_{i1} at each x_i . In this special case let R_i be the rank of Y_{i1} among the set $Y_{i1}, Z_{i1}, Z_{i2}, \dots, Z_{im_i}$. Then, under the null hypothesis, the distribution of R_i is uniform on $1, 2, \dots, m_i + 1$. It is then possible to develop simple nonparametric tests using R_1, R_2, \dots, R_n (see Reynolds, Burkhart and Daniels (1981)).

8. OTHER HYPOTHESIS TESTING APPROACHES

There is a potential problem with testing H_0 and H_0' as previously formulated. It may be known a priori that the model and the real system can not be identical and thus testing that the two are identical may not be very helpful. A more realistic philosophy is to realize that an imperfect model can still be useful and then try to determine how "close" the model needs to be to the real system in order for the model to be useful for its intended purpose. Once this is determined the validation data can be used to test the null hypothesis that the model and system are close enough for the intended application of the model (see, for example, Balci and Sargent (1981)). This approach requires that a measure, say $\lambda(\underline{x})$, of the closeness of F and G be developed. For example, this measure could be $\lambda(\underline{x}) = E(Y - Z|\underline{x})$, the expected difference between the real system output and the model output. The null hypothesis could then be

$$H_0'': \lambda(\underline{x}) \leq \lambda_0 \text{ for all } \underline{x} \in A$$

or, if the required agreement between the real system and model depends on \underline{x} , the null hypothesis could be

$$H_0''': \lambda(\underline{x}) \leq \lambda_0(\underline{x}) \text{ for all } \underline{x} \in A$$

where $\lambda_0(\underline{x})$ is the required agreement at $\underline{X} = \underline{x}$

In order to test H_0'' or H_0''' an appropriate test statistic must be chosen. Balci and Sargent (1981) discuss the use of Hotelling's two-sample T^2 test for this problem when several system response variables are observed and the inferences are not conditional on \underline{x} .

The hypothesis testing approaches discussed so far have all tested the null hypothesis that the model is "valid" in some sense. With this formulation the null hypothesis that the model is valid will be accepted unless there is strong evidence to the contrary. This may lead to the acceptance of a model that is not adequate if the power of the test being used is low. This problem can be overcome somewhat if the power of the test at alternatives of interest can be explicitly controlled.

Another approach that might be more reasonable from the model users point of view is to take the null hypothesis as the hypothesis that the model is not valid. This null hypothesis would then be rejected and the model accepted only if there is strong evidence that the model is valid. In this way the burden of proof is on the model to prove itself before being accepted for use. This approach may be difficult to implement in some cases since the null hypothesis of an invalid model may be difficult to explicitly formulate and test. Reynolds (1984) discusses this approach to formulating the null hypothesis in one particular context.

9. ESTIMATING ERROR

The logical inconsistency in testing the null hypothesis that the model output has the same distribution as the system output when this is known to be impossible has already been pointed out. Testing the hypothesis that the model is close enough for the intended purpose of the model may be more realistic, but there may be problems in implementing this approach. In many cases there will be many potential users of the model. Even if these users can be identified it may be difficult to get these users to accurately specify the required degree of agreement between the model and the real system. In addition, the results of a test may not give the model user much feel for the error that can be expected when the model is used to draw inferences about the real system.

One way around the problems of the hypothesis testing approach is through the approach of what could be called statistical estimation. This approach is concerned with estimating the error that is likely to result when the model is used to estimate a parameter or to predict the actual output of the real system. When the objective is to estimate a parameter then a confidence interval could be given for the difference (expected error) between the mean of the estimator from the model and the actual value of the parameter. When the objective is to predict actual system output in a given situation then a prediction interval for the difference (prediction error) between the prediction and the observed output could be calculated. In this way estimates of error can be used by the model user or users to determine whether the performance of the model is acceptable for various purposes.

The expected output of the system at $\underline{X} = \underline{x}_i$ is $E(Y|\underline{x}_i)$, the expected model output is $E(Z|\underline{x}_i)$, and the expected difference or bias in the model is $E(Y - Z|\underline{x}_i)$. An unbiased estimator of this bias is $D_i = \bar{Y}_i - \bar{Z}_i$. A

confidence interval for this model bias can be constructed to give the model user some indication of the average error that will result when the model is used to estimate the mean response of the system. If m_i and m_i' are not too small then confidence intervals for bias at each point \underline{x}_i can be constructed.

In some cases the objective may be to predict actual system output at some point. If \bar{Z}_i is considered as a predictor of \bar{Y}_i then the prediction error is $D_i = \bar{Y}_i - \bar{Z}_i$. A prediction interval for this error can be constructed to give the model user some indication of the size of the error that may result when the model is used for predicting the response of the system.

If the n sets of conditions can be considered as a random sample from some population then the n values D_1, D_2, \dots, D_n can be used to construct a confidence interval for the average bias (averaged over the distribution of \underline{X}) or to construct a prediction interval for the prediction error at a randomly selected value of \underline{X} . Reynolds (1984) discusses the use of confidence interval and prediction intervals in validating models.

10. REGRESSION

In most cases the difference between the model and the real system will not be constant but instead will vary depending on the values of the input variables. This means that the bias in the model and the distribution of the prediction error will depend on \underline{X} . In addition the accuracy required of the model may also depend on \underline{X} . For example, for certain values of \underline{X} the value of Y may be large and the acceptable error may also be relatively large. But for other values of \underline{X} the value of Y may be small and the acceptable error may also be relatively small. Thus it would be useful to be able to directly relate the error or bias in the model to the levels of the input variables

X. One reasonable approach to this problem is to use regression methodology to relate the error D to the input variables X. If this can be done then model users can obtain information about model accuracy for different conditions. In this case estimates of bias or prediction error would not be restricted to the n validation data points although the regression model for error as a function of X would presumably only be valid within the region of the validation data. Reynolds and Chung (1985) discuss the use of regression methodology in validating models and give an example of this methodology applied to the stand simulator PTAEDA.

11. ACKNOWLEDGEMENT

This research was supported in part by Cooperative Agreement No. 4723-1 with the U.S. Department of Agriculture, Forest Service.

12. REFERENCES

- Balci, Osman and Robert G. Sargent (1981). A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models. Communications of the ACM 24: 190-197.
- Balci, Osman and Robert G. Sargent (1984). A Bibliography on the Credibility Assessment and Validation of Simulation and Mathematical Models. Simuletter 15: 15-27.
- Daniels, Richard F. and Harold E. Burkhart (1975). Simulation of Individual Tree Growth and Stand Development in Managed Loblolly Pine Plantations. FWS-5-75, School of Forestry and Wildlife Resources, Virginia Tech.
- Fisher, R.A. (1938). Statistical Methods for Research Workers. 7th ed. Oliver and Boyd, Edinburgh and London. 356 p.
- Fishman, G.S., and P.J. Kiviat (1968). The Statistics of Discrete-event Simulation. Simulation 10: 185-195.
- Law, Averill M. and W. David Kelton (1982). Simulation Modeling and Analysis. McGraw-Hill, New York.
- McKenney, J.L. (1967). Critique of: "Verification of Computer Simulation Models." Management Science 14:B-102-103.
- Naylor, T.H., and J.M. Finger. (1967). Verification of Computer Simulation Models. Management Science 14:B-92-101.

- Oosterhoff, J. 1969. Combination of One-sided Statistical Tests.
Mathematisch Centrum, Amsterdam, Mathematical Centre Tract 28, 148 p.
- Reynolds, Marion R., Jr. (1984). Estimating the Error in Model Predictions.
Forest Science 30: 454-469.
- Reynolds, Marion R., Jr., Harold E. Burkhardt, and Richard F. Daniels (1981).
Procedures for Statistical Validation of Stochastic Simulation Models.
Forest Science 27: 349-364.
- Reynolds, Marion R., Jr., and Jain Chung (1985). Regression Methodology for
Estimating Model Prediction Error. Submitted for publication.
- Sargent, Robert G. (1979). Validation of Simulation Models. Proceedings of
the Winter Simulation Conference, San Diego, CA, 497-503.
- Sargent, Robert G. (1983). Validating Simulation Models. Proceedings of the
1983 Winter Simulation Conference, edited by S. Roberts, J. Banks,
and B. Schmeiser, 333-337.
- Schrank, W.E. and C.C. Holt (1967). Critique of: "Verification of Computer
Simulation Models." Management Science 14:B-104-106.
- Shannon, Robert E. (1975). Systems Simulation: The Art and the Science.
Prentice Hall.
- Van Horn, R.L. (1971). Validation of Simulation Results. Management Science
17: 247-258.

DISTRIBUTION UNDER DEPENDENCE OF NONPARAMETRIC TWO-SAMPLE TESTS

Emanuel Parzen
Department of Statistics
Texas A&M University

ABSTRACT. This paper aims to show how to develop the theory of two-sample statistical procedures in a way that enables statisticians to determine (in a practical and effective way) how tests can be adjusted for dependence in the case that dependence is modelled by a stationary time series. The importance of the problem of adjusting two-sample tests for dependence is illustrated by an example from Box, Hunter, and Hunter (1978). The paper concludes with a formula for dependence factors of linear rank statistics which are expressed in terms of spectral densities at zero frequency of suitable rank transformed time series. To derive dependence factors, we use the asymptotic distribution theory of sample distribution functions and sample quantile functions of stationary time series. Proofs of these results and examples of their applications are given by A. Harpaz (1985) in his Ph.D. thesis.

1. INTRODUCTION

Serial dependence (autocorrelation) in data can seriously affect the performance of standard statistical procedures (such as the t-test or Wilcoxon rank sum test for the equality of location parameters of two samples). The qualitative truth of this statement is well known to statisticians. But general techniques for evaluating quantitatively the properties of standard statistical procedures under dependence are not being used by statisticians. This paper aims to show how to develop the theory of two-sample statistical procedures in a way that enables statisticians to determine (in a practical and effective way) dependence factors which adjust tests in the case that dependence is modelled by a stationary time series.

To illustrate and motivate the importance of the problem of adjusting two-sample tests for dependence we quote an example presented by Box, Hunter, and Hunter (1978, pp. 81-82). An experiment is performed which takes two samples of 10 observations each from identical populations and tests for a change in location by a t test and a Wilcoxon test using a 5% level of significance. This experiment was repeated 1000 times and one observed the percentage P of the number of experiments in which the null hypothesis of equality of distributions is rejected. When the samples of size 10 consist of independent observations one expects that, and observes that, approximately

Research supported by the U. S. Army Research Office Grant DAAG29-83-K-0051.

$P=5\%$. The experiment also simulated observations with errors $e(t)$ generated from white noise $u(t)$ by a first order moving average model $e(t)=u(t)+bu(t-1)$, with b chosen so that the lag one autocorrelation ρ equaled $-.4$ (negative autocorrelation) or $.4$ (positive autocorrelation). Under these conditions the values observed for P were very approximately $P=11\%$ for $\rho=.4$ and $P=0.2\%$ for $\rho=-.4$. One would like to be able to compute theoretical values of P which can be compared with, and help us understand and predict, the observed values of P . The formulas given in this paper show that the theoretical values of P depend in large samples on the value, denoted $f(0)$, at zero frequency of the spectral density function of the time series model describing the dependence of the observations.

For a first order moving average $f(0) = 1+2*\rho$, so that $f(0)=1.8$ for $\rho=.4$ and $f(0)=.2$ for $\rho=-.4$; note that $f(0)=1$ for white noise ($\rho=0$). These values of $f(0)$ can be used to compute theoretical values of P (based on sampling theory for dependent data) which are in rough accord with the values of P observed by Box, Hunter, and Hunter in their experiment. The conclusion drawn by Box, Hunter, and Hunter from their experiment is that the significance levels of the t and Wilcoxon tests are affected remarkably little by dramatic changes in the probability distribution (normal, uniform, skewed) but are seriously impaired by serial dependence. To resolve the problem of dependent errors one approach is to avoid dependence through randomization. But when serial dependence cannot be avoided its effect must be assessed quantitatively. This paper describes methods for adjusting (for time series dependence) two-sample linear rank tests to have known sampling distribution under the null hypothesis.

As an example, let us note that the z -statistics in eq. (3.29) or the t -statistic in eq. (3.33) of Box, Hunter, and Hunter (1978) could be approximately adjusted for serial dependence by dividing by $\{f(0)\}^{1/2}$. This formula generalizes the discussion on p. 588 of Box, Hunter, and Hunter (1978). [When $f(0) = .2$, its square root is $.45$. The adjusted t -statistic $1.01/.45 = 2.26$ or adjusted t -statistic $.88/.45 = 1.96$ yield P -levels comparable to that of the t -value 2.17 obtained in eq. (2.16)].

2. LINEAR RANK STATISTICS DEPENDENCE FACTORS

Let $X(1), \dots, X(m)$ be a sample from a strictly stationary time series with distribution function $F(x) = \text{PROB}[X \leq x]$, $-\infty < x < \infty$, and quantile function

$$Q(u) = F^{-1}(u) = \inf \{x: F(x) \geq u\}, \quad 0 < u \leq 1.$$

The population mean and variance of X are denoted MX and $VARX$. The sample mean and variance of $X(1), \dots, X(m)$ are denoted $MX\{m\}$ and $VARX\{m\}$.

Let $Y(1), \dots, Y(n)$ be a sample from a strictly stationary time series with distribution function $G(x) = \text{PROB}[Y \leq x]$, $-\infty < x < \infty$ and quantile function $G^{-1}(u)$. Assume that X values are independently distributed from Y values.

Let T denote a linear rank statistic to test the null hypothesis H_0 of equality of the distributions $F(x)$ and $G(x)$. To compute and represent T one introduces the rank, denoted R_j , of the j -th largest X value within the pooled sample of X and Y values. A typical definition of T is

$$(1) \quad T = (1/m) \sum_{j=1}^m J(R_j / (N+1))$$

where $N=m+n$ is the pooled sample size and $J(u)$, $0 \leq u \leq 1$, is a suitable score function. The Wilcoxon rank-sum test corresponds to $J(u)=u$ or $J(u)=u-0.5$.

The asymptotic distribution of T under the null hypothesis H_0 can be described in terms of $\lambda=m/N$, $MJ(U) = \int_0^1 J(u) du$, and

$$\text{VARJ}(U) = \int_0^1 \{J(u) - MJ(U)\}^2 du.$$

The role of U will become clear in the sequel (section 4); it represents a random variable with a uniform distribution on the interval 0 to 1. This paper shows how to express the asymptotic distribution of T , as N tends to ∞ , in the form

$$\sqrt{N}(T - MJ(U)) \text{ is NORMAL}(0, ((1-\lambda)/\lambda) * \text{VARJ}(U) * \text{DEPFAC}[T])$$

The notation $*$ denotes multiplication

We use $\text{DEPFAC}[T]$ to denote dependence factor of T ; it equals 1 if the X 's are independent random variables and Y 's are independent random variables. The main aim of this paper is to present a formula for the dependence factor $\text{DEPFAC}[T]$ of a linear rank statistic T . To adjust T for dependence we could use $(T - MJ(U)) / \{\text{DEPFAC}[T]\}^{1/2}$ as our test statistic.

To help interpret and understand the formula we present at the end of the paper for $\text{DEPFAC}[T]$ the next section introduces dependence factors for sample means.

3. DEPENDENCE FACTORS AND SPECTRAL DENSITIES AT ZERO FREQUENCY

Our notation for the theoretical mean and variance of a random variable X is $MX=E[X]$ and $\text{VARX} = E\{(X-MX)^2\}$. When $X(t)$, $t=0, \pm 1, \pm 2, \dots$, is a stationary time series its covariance function is denoted $R(v; X) = \text{COV}[X(t), X(t+v)]$ and its correlation function is denoted

$$\text{RHO}(v;X) = R(v;X)/R(0;X) = \text{CORR}[X(t), X(t+v)], \\ v=0, \pm 1, \pm 2, \dots$$

The sample mean of $X(1), \dots, X(n)$ is denoted

$$\text{MX}(n) = (1/n) \sum_{t=1}^n X(t)$$

The variance of a sample mean can be expressed

$$n \text{VAR}[\text{MX}(n)] = \text{VARX} * \text{DEPFAC}[\text{MX}(n)]$$

where

$$\text{DEPFAC}[\text{MX}(n)] = \sum_{v=-n}^n (1 - |v/n|) \text{RHO}(v;X)$$

In words, the variance of the sample mean of a stationary time series can be represented as the product of its variance for an independent sample and a dependence factor.

For large samples (as n tends to ∞) one can relate the dependence factor to the spectral density of the time series, denoted

$$\text{SPECDEN}(\omega;X) = 1 + 2 \sum_{v=1}^{\infty} \text{RHO}(v;X) \cos 2\pi\omega v, \quad 0 \leq \omega \leq 1.$$

For n large, the dependence factor of a sample mean is given by

$$\text{DEPFAC}[\text{MX}(n)] = \text{SPECDEN}(0;X)$$

The advantage of expressing the dependence factor in terms of the spectral density at zero frequency is that it can be estimated using methods of spectral density estimation.

Let us now consider the two-sample problem of testing the equality of distributions of two independent time series $X(t)$ and $Y(t)$ using as a test statistic the difference of the sample means

$$\text{MX}(m) = (1/m) \sum_{t=1}^m X(t), \quad \text{MY}(n) = (1/n) \sum_{t=1}^n Y(t).$$

The test statistic $\text{MX}(m) - \text{MY}(n)$ has variance equal to the sum of the variances of the two sample means. Therefore approximately

$$\text{VAR}[\text{MX}(m) - \text{MY}(n)] = (1/m) \text{VARX} * \text{SPECDEN}(0;X) + (1/n) \text{VARY} * \text{SPECDEN}(0;Y)$$

Assume that under H_0 both $MX=MY$ and $VARX=VARY$ (in practice, one might replace $VARX$ and $VARY$ by the variance of the pooled sample). Then under H_0 $MX\{m\}-MY\{n\}$ has mean 0 and variance

$$(1) \quad \text{VAR}[MX\{m\}-MY\{n\}] = \{N\lambda(1-\lambda)\}^{-1} * \text{VARX} * \text{DEPFAC}[MX\{m\}-MY\{n\}]$$

where $N=m+n$, $\lambda=m/N$, and the dependence factor can be expressed approximately (for large values of m and n) in terms of spectral densities:

$$(2) \quad \text{DEPFAC}[MX\{m\}-MY\{n\}] = (1-\lambda) \text{SPECDEN}(0;X) + \lambda \text{SPECDEN}(0;Y)$$

It should be noted that we are not assuming that the spectral densities of X and Y are equal.

This formula for the dependence factor of the difference of two means is important for several reasons:

(1) It can be used to determine the affect of dependence on the two sample t-test; it shows that the affect for large samples depends only on the value of the spectral densities of $X(t)$ and $Y(t)$ at zero frquency.

(2) It motivates the form of answer which we seek for linear rank statistics T , since we shall show that $T - MJ(U)$ has the same distribution as a difference-of-means statistic

$$(1-\lambda) (MJ(UX^{\sim})\{m\} - MJ(UY^{\sim})\{n\})$$

in terms of time series $J(UX^{\sim}(t))$ and $J(UY^{\sim}(t))$ defined below. The asymptotic variance of T therefore can be expressed, using (1) and (2),

$$\frac{(1-\lambda)^2}{N\lambda(1-\lambda)} * \text{VARJ}(UX) * \{(1-\lambda) \text{SPECDEN}(0;J(UX^{\sim})) + \lambda \text{SPECDEN}(0;J(UY^{\sim}))\}$$

The remarkable conclusion which one is able to draw from this formula is that for large samples the dependence factor of linear rank statistics can be evaluated by estimating the spectral density at zero frequency of the derived time series $J(UX^{\sim}(t))$ and $J(UY^{\sim}(t))$. Experience indicates that a quick and dirty estimate of these spectral densities is provided by the spectral densities of $X(t)$ and $Y(t)$ respectively. In practice one will not know the dependence structure of the errors. The dependence factor of T will be estimated by estimating the spectral density at zero frequency of the time series whose means are being compared.

4. REPRESENTATIONS OF LINEAR RANK STATISTICS

To study linear rank statistics we use representations for them in terms of sample distribution functions which are valid for both independent and dependent observations.

A sample $X(1), \dots, X(m)$ has: order statistics $X(1;m) \leq \dots \leq X(m;m)$; sample distribution function $F^{\sim}(x)$ = fraction of sample $\leq x$; and sample quantile function $Q^{\sim}(u) = F^{\sim-1}(u)$ given by

$$Q^{\sim}(u) = X(j;m) \text{ for } (j-1)/m < u \leq j/m.$$

One also uses continuous versions of the discrete sample quantile function. A sample $Y(1), \dots, Y(n)$ has: order statistics $Y(1;n) \leq \dots \leq Y(n;n)$ and sample distribution function $G^{\sim}(x)$.

One pools the two samples to form a pooled sample $X(1), \dots, X(m), Y(1), \dots, Y(n)$ of size $N=m+n$ which has sample distribution function $H^{\sim}(x)$ satisfying $H^{\sim}(x) = \lambda F^{\sim}(x) + (1-\lambda)G^{\sim}(x)$. The limit of $H^{\sim}(x)$ is $H(x) = \lambda F(x) + (1-\lambda)G(x)$.

In the one-sample problem we call $U(t) = F(X(t))$, $t=1, \dots, m$, the rank transformed variables; their marginal distribution is uniform on 0 to 1. Sample rank transformed variables $U^{\sim}(t)$ are defined by a formula such as $U^{\sim}(t) = (m/(m+1))F^{\sim}(X(t))$ which assigns ranks $1/(m+1), \dots, m/(m+1)$ to the order statistics $X(1;m), \dots, X(m;m)$.

In the two-sample problem the rank transformed variables are defined to be $H(X(t))$ and $H(Y(t))$. The sample rank transformed variables are

$$UX^{\sim}(t) = (N/(N+1))H^{\sim}(X(t)), \quad UY^{\sim}(t) = (N/(N+1))H^{\sim}(Y(t)).$$

A linear rank statistic T as traditionally defined by eq (1) of section 2 can be represented

$$T = (1/m) \sum_{j=1}^m J\left(\frac{N}{N+1} H^{\sim}X(j;m)\right) = MJ(UX^{\sim})\{m\}.$$

An alternative statistic, which our analysis shows provides more insight into the asymptotic distribution, is the difference-of-means statistic; one can show that asymptotically [and exactly for $J(u)=u$]

$$T - MJ(U) = (1-\lambda)(MJ(UX^{\sim})\{m\} - MJ(UY^{\sim})\{n\})$$

To relate T to sample distribution functions we represent it

$$T = \int_{-\infty}^{\infty} J\left(\frac{N}{N+1} H^{\sim}(x)\right) dF^{\sim}(x)$$

Our approach is to write approximately

$$T = \int_0^1 J(u) dF^{\sim}(H^{\sim-1}(u)).$$

This formula is not used. But it suggests one should try to represent T exactly as

$$T = \int_0^1 J(u) d\tilde{D}(u)$$

where $\tilde{D}(u)$ is a suitable estimator of $D(u) = F^{-1}(u)$, $0 \leq u \leq 1$. We call $D(u)$ a comparison quantile function.

We would like to define $\tilde{D}(u)$ in terms of sample distribution functions so that it is a step function with jumps equal to $1/m$ at $u = (N/(N+1)) R_j$. Parzen (1983) shows that this can be accomplished if $\tilde{D}(u)$ is defined as the inverse $D_1^{-1}(t)$ of $D_1(t) = H\tilde{F}^{-1}(t)$, $0 \leq t \leq 1$.

Our motivations for introducing $D(u)$ and $\tilde{D}(u)$ are diverse.

(1) They implement our philosophy that every graph should be a picture of a function. Various techniques for graphical analysis of samples, such as P-P plots and Q-Q plots, can be regarded as sample versions of theoretical functions of the form of $D(u)$.

(2) The conclusions that one obtains arithmetically from the value of a linear rank statistic can often be discovered graphically (at a glance) from a graph of $\tilde{D}(u)$.

(3) In cases where the value of T indicates no significant difference between the two samples, the graph of $\tilde{D}(u)$ may indicate important ways in which the samples differ.

(4) The empirical process $\tilde{D}(u)$ is important as a practical basis for data analysis (as outlined in reasons (2) and (3)) and as a theoretical basis for deriving the properties of linear rank statistics. The asymptotic distribution of $\tilde{D}(u)$, $0 \leq u \leq 1$, is derived by expressing it in terms of the asymptotic distributions of the sample distribution functions of the independent stationary time series $X(t)$ and $Y(t)$. The rigorous theory of the latter has recently been completed by Pham and Tran (1985) as the culmination of a long line of research papers starting with the pioneering work of Gastwirth and Rubin (1975).

5. EMPIRICAL PROCESSES OF STATIONARY TIME SERIES

Let $\tilde{F}(x)$ and $\tilde{Q}(u)$ denote the sample distribution and sample quantile function of $X(1), \dots, X(n)$, a sample from a stationary time series $X(t)$. Let $CFX(x)$, $-\infty < x < \infty$, and $CF^{-1}X(u)$, $0 \leq u \leq 1$, denote stochastic processes representing the limiting distributions of $\sqrt{n}\{\tilde{F}(x) - F(x)\}$, $-\infty < x < \infty$, and $\sqrt{n}\{\tilde{F}^{-1}(u) - F^{-1}(u)\}$, $0 \leq u \leq 1$, respectively. One can show that there is a zero mean Gaussian stochastic process denoted $BX(u)$, $0 \leq u \leq 1$, such that

$$CFX(x) = BX(F(x)), \quad CF^{-1}(u) = \{-1/fF^{-1}(u)\} BX(u).$$

Thus the asymptotic distribution of the sample distribution and sample quantile functions can be expressed in terms of the process $BX(u)$, $0 \leq u \leq 1$.

For independent random variables (white noise) $X(t)$, the limit process $BX(u)$ is a Brownian Bridge, which is a zero mean Gaussian process with covariance kernel

$$E[BX(u_1) BX(u_2)] = u_1 (1-u_2) \quad \text{for } u_1 \leq u_2.$$

An indication of the formulas required to describe $BX(u)$ when $X(t)$ is a time series is provided by the limit distributions of independent samples of bivariate dependent random variables ($X(t)$, $Y(t)$). Then the limit processes $BX(u)$ and $BY(u)$ are each Brownian Bridges but they are not independent of each other. They have joint covariance kernel

$$E[BX(u_1) BY(u_2)] = F(QX(u_1), QY(u_2)) - u_1 u_2, \quad 0 \leq u_1, u_2 \leq 1,$$

where $F(x, y) = \text{PROB}[X \leq x, Y \leq y]$ is the joint distribution function of X and Y . We call $F(QX(u_1), QY(u_2))$ the bivariate dependence function of X and Y ; an alternative name (used by some authors) is copula.

To express the covariance kernel of $BX(u)$, $0 \leq u \leq 1$, in the case that $X(t)$ is a stationary time series, it is more convenient (for insight and computation and to avoid a complicated infinite summation of bivariate dependence functions) to represent the covariance structure as a formula for the variance of a general linear functional $\int_0^1 g(u) dBX(u)$ for suitable functions $g(u)$. Let $U(t) = F(X(t))$ be the rank transform, and form the time series $g(U)$ whose value at t is $g(U(t))$. Equivalently we write $g(U) = g(F(X))$.

BASIC THEOREM ON EMPIRICAL PROCESS OF STATIONARY TIME SERIES: The distribution of $BX(u)$, $0 \leq u \leq 1$, can be described in terms of the spectral density at zero frequency of the time series $gU(t)$, $U(t) = F(X(t))$, which are estimated by $g\tilde{U}(t)$, $\tilde{U}(t) = \tilde{F}(X(t))$:

$$\text{VAR}[\int_0^1 g(u) dBX(u)] = \text{VAR}g(U) \text{SPECDEN}(0; g(U))$$

where

$$\text{VAR}g(U) = \int_0^1 g^2(u) du - |\int_0^1 g(u) du|^2.$$

The asymptotic distribution of linear rank statistics are obtained from formulas for the asymptotic distribution of linear

functionals in the sample comparison quantile function $\tilde{D}(u)$, $0 < u < 1$, defined in section 4. One can show that (in the sense of convergence of stochastic processes)

$$\sqrt{N} \{ \tilde{D}(u) - D(u) \} \rightarrow CD(u)$$

where the limit process $CD(u)$, $0 < u < 1$, can be expressed in terms of independent limit processes $BX(u)$, $0 < u < 1$, and $BY(u)$, $0 < u < 1$, by

$$CD(u) = -(1-\lambda) \{ \lambda^{-1/2} BX(u) - (1-\lambda)^{-1/2} BY(u) \}$$

The processes $BX(u)$ and $BY(u)$ are related to the processes defined in the Basic Theorem on Empirical processes. Their covariance kernels are expressed in terms of the spectral densities at zero frequency of the time series $J(UX^{\sim}(t))$ and $J(UY^{\sim}(t))$:

$$\text{VAR} \left[\int_0^1 J(u) dBX(u) \right] = \text{VAR} J(U) \text{SPECDEN}(0; J(UX^{\sim})),$$

$$\text{VAR} \left[\int_0^1 J(u) dBY(u) \right] = \text{VAR} J(U) \text{SPECDEN}(0; J(UY^{\sim})).$$

By combining all these results one can obtain the formula given in section 2 for the asymptotic distribution of a linear rank statistic for two samples from stationary time series with dependence factor $\text{DEPFAC}[T]$ estimated by

$$\text{DEPFAC}[T] = (1-\lambda) \text{SPECDEN}(0; J(UX^{\sim})) + \lambda \text{SPECDEN}(0; J(UY^{\sim}))$$

A more complete proof of this result, and examples of its applications, are given by Harpaz (1985) in his Ph.D. thesis.

REFERENCES

- Box, G.E.P., Hunter, W.G., and Hunter, J. S. (1978) Statistics for Experimenters, Wiley: New York.
- Gastwirth, J. L. and Rubin, H. (1975) "The Asymptotic Distribution Theory of the Empirical c.d.f. for Mixing Stochastic Processes," Ann. Statistics, 3, 809-824.
- Harpaz, A. (1985) "Stationary Time Series, Quantile Functions, Nonparametric Inference, and Rank Transform Spectrum." Technical Report A-31. Statistics Department, Texas A&M University.
- Parzen, E. (1983) "Fun.Stat Quantile Approach to Two Sample Statistical Data Analysis," Technical Report A-21, Statistics Department, Texas A&M University.
- Pham, Tuan D. and Tran, Lanh T. (1985) "Some Mixing Properties of Time Series Models," Stochastic Processes and their Applications, 19, 297-303.

Statistical Models and Methods for
CLUSTER ANALYSIS AND SEGMENTATION

Stanley L. Sclove
Department of Information and Decision Sciences
College of Business Administration
University of Illinois at Chicago

ABSTRACT

Clustering of individuals, segmentation of time series and segmentation of numerical images can all be considered as labeling problems, for each can be described in terms of pairs (x_t, g_t) , $t = 1, 2, \dots, n$, where x_t is the observation at instance t and g_t is the unobservable "label" of instance t . The labels are to be estimated, along with any unspecified distributional parameters. In cluster analysis the values of t are the individuals (cases) observed and the x 's are independent. In time series the values of t are time instants and there is temporal correlation. In numerical image segmentation the values of t denote picture elements (pixels) and spatial correlation between neighboring pixels can be utilized. The idea in segmentation is that signals and time series often are not homogeneous but rather are generated by mechanisms or processes with various phases. Similarly, images are not homogeneous but contain various objects. "Segmentation" is a process of attempting to recover automatically the phases or objects. A labeling model for representing such signals, time series, and images was discussed in a paper by the present author in the Proceedings of the 30th Conference; some approaches to estimation and segmentation in this model were presented. The present paper summarizes the work on all these types of labeling problems, clustering as well as time series- and image-segmentation.

Key words and phrases: statistical pattern recognition, classification; temporal correlation, spatial correlation; optimization by relaxation method.

1. Introduction

The research reported here relates to cluster analysis and numerical processing of time series and images. It is in part a discussion of work performed under ARO Contract DAAG29-82-K-0155 (6/15/82 - 6/15/85): Statistical Models and Methods for Cluster Analysis and Image Segmentation. The type of datasets to which the techniques developed are applicable include: signals such as radar and sonar; economic and bio-medical time series; time series arising from quality assurance acceptance sampling by attributes or variables; and digital images which can result from various sources, including bio-medical imagery, infrared imagery obtained by smart munitions, and multispectral data obtained by satellite. The problems addressed are those of clustering, and segmentation of time series and images.

The work involves the further development of algorithms for clustering large, multidimensional datasets and for segmentation of time series and digital images. The algorithms are based on maximum likelihood estimation in distribution-mixture models. In the context of these mixture models clustering is construed as estimation of unobserved labels. An observation's label, were it observable, would tell from which mixture component the observation arose. Image segmentation is also considered as a labeling problem. Throughout the work there is an attempt to apply model-selection criteria to the decision as to an appropriate number of clusters or classes of segment.

Software development is an important aspect of such a project. The algorithms developed are programmed in FORTRAN.

Some of the ideas discussed in the present paper have been

developed and published in journals; see Sclove (1977; 1983a,b,c; 1984a) and Bozdogan and Sclove (1984).

The organization of the present paper is as follows: Section 2 concerns cluster analysis; in this section there is some general discussion of model-selection criteria and a digression to mention some ideas concerning clustering of variables. Section 3 summarizes some of the results on time-series segmentation, and results on image segmentation are discussed in Section 4.

2. Cluster analysis

Background. The mixture model for the clustering problem postulates a mixture of k distributions. This is the approach put forth in (Sclove 1977). The research problem set there was, at least in part, to see whether the ISODATA (Ball and Hall, 1967) and K-MEANS (MacQueen, 1967) algorithms could be interpreted as mathematical-statistical estimation schemes in some model for the clustering problem. That is, did there exist a model for the clustering problem, and an estimation method in that model, such that ISODATA and K-MEANS corresponded to that method applied to that model? The answer, provided in (Sclove 1977), was affirmative; this will be explained below, but first let us briefly define ISODATA and K-MEANS.

The "isodata" scheme proceeds as follows. One starts with tentative estimates of cluster means as seed points for the clusters and assigns each observation to the mean to which it is closest. The cluster means are then re-estimated, and one loops through the data again, reassigning the observations. Etc. In the K-MEANS algorithm, the seed points are updated immediately after each observation is

tentatively classified. In (Sclove 1977) it was shown that these algorithms correspond to iterative maximum likelihood estimation in a type of mixture model for the clustering problem, where the component distributions are multivariate normal.

This clustering can be done for various values of k , the number of clusters. Figures of merit can be used to choose the best k . Model-selection criteria can be used as figures of merit.

2.1. Model-selection criteria

In the context of a mixture model, choice of the number of clusters k can be viewed as a model-selection problem. However, at least in the case of clustering individuals, existing model-selection criteria have to be modified, as they depend upon (regularity) assumptions that are not always met in mixture models for clustering individuals.

In any case, let us review some of the existing model-selection criteria. Consider, then, a problem of choosing from among several models, indexed by k ($k = 1, 2, \dots, K$). Let $L(k)$ be the likelihood, given the k -th model. Various model-selection criteria taking the form

$$-2 \log(\max L(k)) + a(n)m(k) + b(k), \quad (1)$$

have been developed in relatively recent years. Here n is the sample size, \log denotes the natural logarithm, $\max L(k)$ denotes the maximum of the likelihood over the parameters, and $m(k)$ is the number of independent parameters in the k -th model. For a given criterion, $a(n)$ is the cost of fitting an additional parameter and $b(k)$ is an additional term depending upon the criterion and the model k .

Akaike (see, e.g., Akaike 1973, 1974, 1981) developed such a

criterion as an (heuristic) estimate of the expected entropy (Kullback-Leibler information). Akaike's information criterion (AIC) is of the form (1) with

$$a(n) = 2 \text{ for all } n, \quad b(k) = 0 \quad (\text{AIC}). \quad (2)$$

Schwarz (1978), working from a Bayesian viewpoint, obtained a criterion of the form (1) with

$$a(n) = \log n, \quad b(k) = 0 \quad (\text{Schwarz' criterion}). \quad (3)$$

Since, for n greater than 8, $\log n$ exceeds 2, it follows that Schwarz' criterion favors models with fewer parameters than does Akaike's.

Noting that AIC has $a(n)$ a constant function of n , namely 2, various researchers, including Kashyap (1982) and Schwarz (1978) have mentioned that AIC is not consistent; $a(n)$ needs to depend upon n .

Kashyap (1982), also working from a Bayesian approach, took the asymptotic expansion of the logarithm of the posterior probabilities a term further than did Schwarz and obtained the criterion of the form (1) given by

$$a(n) = \log n, \quad b(k) = \log(\det B(k)) \quad (\text{Kashyap's criterion}), \quad (4)$$

where \det denotes the determinant and $B(k)$ is the negative of the matrix of second partials of $\log L(k)$, evaluated at the maximum likelihood estimates. In Gaussian linear models this is the covariance matrix of the maximum likelihood estimates of the regression coefficients; in general, the expectation of $B(k)$, evaluated at the true parameter values, is Fisher's information matrix. Since Kashyap's criterion is based on reasoning similar to Schwarz', but contains an extra term, it may perform better. [Further comments on model-selection criteria are made in Sclove (1983d).]

2.2. Multi-sample clustering

The problem of multi-sample clustering, the grouping of samples, is treated in Bozdogan and Sclove (1984). The situation is the K-sample problem (one-way analysis of variance), with an emphasis on grouping the samples into fewer than K clusters. The use of model-selection criteria in this context can provide an alternative to multiple-comparison procedures. Use of model-selection criteria avoids the difficult choice of levels of significance in such problems. Model-selection criteria can also be used in this context to decide whether or not to assume a common covariance matrix. Kashyap's criterion could be evaluated and used for these problems.

2.3. Clustering of individuals

Schwarz' and Kashyap's criteria could be calculated for the problem of clustering individuals according to Wolfe's (1970) mixture-model clustering approach and incorporated into computer programs for clustering. The values of the criteria can be used heuristically as figures of merit for alternative models, but in order to be rigorously applied the model-selection criteria need to be modified since their derivation involves an assumption of nonsingularity of the information matrix. However, note in this regard a potential advantage of model-selection criteria over a hypothesis-testing approach in this and similar situations. Model-selection criteria require nonsingularity of the information matrix only for each fixed model k . The testing approach runs into difficulties because of nonsingularity of the matrix at the boundary between the null and alternative hypotheses (i.e., at the boundary between models).

2.4. Clustering of variables

The clustering of variables can also be viewed as a model-selection problem. For example, whether and how to cluster multinormal variables depends upon which covariances may be assumed to be zero; the possible patterns of zeros among the covariances are separate models, a figure of merit for which is provided by a suitable model-selection criterion. This idea is to be further developed.

3. Time-series segmentation

As mentioned above, a model for clustering or segmentation is given by assuming that each instance of observation, t , gives rise not only to an observation x_t but also to a label, g_t , equal to 1, 2, ..., or k , where k is the number of classes of segment. Model-selection criteria are used to estimate k . In the context of this model, segmentation is merely estimation of the labels. Sclove (1983b,c; 1984a) treats the problem by modeling the label process as a Markov chain. An algorithm and computer programs are discussed; numerical examples are given.

The model involves three sets of parameters: the distributional parameters (e.g., means and covariance matrices), the labels, and the transition probabilities between labels.

The algorithm is a relaxation method, similar to the EM algorithm. The estimation step consists of maximum-likelihood estimation of the distributional parameters, for tentatively fixed values of the labels and transition probabilities. The maximization step consists of maximizing the likelihood over the labels and transition probabilities, for tentatively fixed values of the distributional parameters.

As developed so far, the algorithm is a forward algorithm, classifying x_2 after x_1 , x_3 after x_2 and x_1 , etc. It is suitable for sequential operation in real time, but it is non-optimal in other modes of operation. Its performance could possibly be improved by a backcasting technique analogous to that in Box and Jenkins (1976) and by application of the Viterbi algorithm (Forney 1973), which is a recursive optimal solution to the problem of estimating the state sequence of a discrete-time finite state Markov process; it is applicable here because this is what we have at each stage when the distributional parameters and transition probabilities are tentatively fixed and the labels are to be estimated.

Further, the parameter-estimation step of the algorithm can be improved. The estimation implemented in the existing algorithm leads to estimates that are biased (even asymptotically). (See, e.g., Bryant and Williamson 1978.) This bias may be viewed as due to the truncation resulting from the algorithm. The estimation could be modified by doing it in a Bayesian manner, e.g., estimate the mean of Class A as

$$\frac{\sum_{t=1}^n x(t) \Pr(a|x(t))}{\sum_{t=1}^n \Pr(a|x(t))}$$

(In this expression, $\Pr(a|x)$ can be replaced by $\Pr(x|a)$ since $\Pr(a)/f(x)$ will cancel out.) This modification in the parameter-estimation step can be important. For, in this estimate, all the observations play a role, whether labeled as "Class A" or otherwise, so that at least some of the bias incurred by using only the "a" observations will be removed by allowing all of the observations to enter.

The work done to date is explicit only for the case in which the class-conditional processes consist of independent, identically distributed random variables. The work is to be extended to other, often more realistic cases, such as that of autoregression within segments.

4. Image segmentation

Similar ideas are applied to digital images in Sclove (1983a;1984a). Here the label process is modeled as a Markov random field. The same improvements made in the time-series context will be carried over to the two-dimensional, image-processing context. For example, computer experiments (Sclove 1984b) with the existing algorithm have shown it to be successful, even in finding small targets. However, at the same time, these experiments have shown the importance of some such modification as backcasting, as mentioned in connection with time series, to eliminate anomalous border effects.

Extension of the existing work to two-dimensional autoregressions within segments will yield algorithms that may detect textures.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Proceedings of the 2nd International Symposium on Information Theory, 267-281. Akademia Kiado, Budapest.
- Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 6, 716-723.
- Akaike, H. (1981). Likelihood of a model and information criteria. Journal of Econometrics 16, 3-14.
- Ball, G. H., and Hall, D. J. (1967). A clustering technique for summarizing multivariate data. Behavioral Science 12, 153-155.
- Box, G.E.P., and Jenkins, G.M. (1976). Time Series Analysis:

- Forecasting and Control, rev. ed. John Wiley & Sons, New York.
- Bozdogan, Hamparsum, and Sclove, Stanley L. (1984). Multi-sample cluster analysis using Akaike's information criterion. *Annals of the Institute of Statistical Mathematics* 36, 163-180.
- Bryant, P., and Williamson, J. A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* 65, 273-281.
- Forney, G. David, Jr. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, Vol. 61, 268-278.
- Kashyap, R. L. (1982). Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4, 99-104.
- MacQueen, J. (1966). Some methods for classification and analysis of multivariate observations. Pages 281-297 in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Los Angeles and Berkeley.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.
- Sclove, Stanley L. (1977). Population mixture models and clustering algorithms. *Communications in Statistics(A)* 6, 417-434.
- Sclove, Stanley L. (1983a). Time-series segmentation: a model and a method. *Information Sciences* 29, 7-25.
- Sclove, Stanley L. (1983b). Application of the conditional population-mixture model to image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5, 428-433.
- Sclove, Stanley L. (1983c). On segmentation of time series. In *Studies in Econometrics, Time Series, and Multivariate Statistics* (S. Karlin, T. Amemiya, and L. Goodman, eds.), Academic Press, 1983, 311-330.
- Sclove, Stanley L. (1983d). Use of model-selection criteria in clustering and segmentation of time series and digital images. Contributed paper, 44th Session of the International Statistical Institute, Madrid, 9/12-22/83.
- Sclove, Stanley L. (1984a). On segmentation of time series and images in the signal detection and remote sensing contexts. Pages 421-434 in *Statistical Signal Processing* (Edw. J. Wegman and James G. Smith, eds.), Marcel Dekker, Inc., New York.
- Sclove, Stanley L. (1984b). On Segmentation of Signals, Time Series, and Images. Pages 267-289 in *Proceedings of the 30th Conference on Design of Experiments in Army Research, Development and Testing*, Las Cruces, NM, 10/15-19/84 (ARO Report 85-2).
- Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 5, 329-350.

A COMPARISON OF METHODS FOR FACTOR ANALYSIS OF VISIBILITY

Oskar M. Essenwanger
Research Directorate
Research, Development, and Engineering Center
U.S. Army Missile Command
Redstone Arsenal, Alabama 35898-5248

ABSTRACT: Visibility is produced by a variety of meteorological factors related to micro-, meso-, and macro-scale processes. In addition the frequency distribution of visibility is non-Gaussian. Thus a factor analysis is not trivial.

Today factor analysis is aided by "canned" programs on most larger computer systems. However, most of the time it is not readily understood what these programs produce. Thus an investigation was performed to compare four different approaches of a factor analysis. A principal components analysis, an unweighted least squares, a general least squares approach and a maximum likelihood method were examined for a basic correlation matrix of eight atmospheric parameters and for a 7-year record of Stuttgart, Germany. Furthermore, unrotated factors, and orthogonal and oblique rotation of factors were included. As expected the results of the factor analysis differ in details. However, the four methods show some common principles.

1. **INTRODUCTION:** Factor analysis was used in behavioral science when Spearman (1904, 1927), Cattell (1952 and 1965), and others established the basic statistical-mathematical background. The physical sciences followed hesitantly. Factor analysis in the atmospheric sciences can only be found in the last two decades, e.g. Christensen and Bryson (1966), Kutzbach (1967), Buell (1971) etc.

In part this was due to the elaborate mathematical procedure which is required in the mathematical solution. Today, factor analysis is aided by electronic data processing. In recent times even "canned programs" are available. Thus the mathematical difficulties have been resolved. The physicist will find several methods of estimation, however, and may be confused about the answer to the question which method may be most suitable and may provide the best estimators. Furthermore, in order to

draw the correct conclusions from the solutions by those "canned programs," it is necessary to separate the "mathematics" from the "physics."

This study serves to elucidate some of the mathematical background and reveal some physical characteristics by comparing the results for several methods of factor analysis applied to data of a seven-year record of atmospheric parameters for Stuttgart, Germany.

We learn that the estimators for the "communalities" differ for the individual methods. This is expected. The physical characteristics of the factors, however, display great similarity after rotation of the coordinate system although the sequence is not always the same for the individual methods.

2. PRINCIPAL COMPONENTS ANALYSIS. The basic model for factor analysis can be formulated as follows:

$$M_X = M_A M_F + M_\epsilon \quad (1)$$

where M_X is a data matrix (the only known matrix in Eqn 1), M_A a coefficient matrix of factors, M_F the factor matrix, and M_ϵ an error matrix. M_A is also called the factor loading matrix or factor pattern. In the basic factor analysis neither the factors are correlated, nor are the factors and the errors.

The mathematical solution for Eqn (1) can be formulated as:

$$M_X = M_A \Phi M_A^T + (\Psi) \quad (2)$$

where $\Phi = M_F^T M_F$ is a factor covariance matrix and Ψ a diagonal matrix $\Psi = M_D M_\epsilon$, with M_D a diagonal error matrix.

As stated above, M_X is a data matrix. In its standardized form M_X is a correlation matrix M_R with unity in its diagonal. This is called a "closed" system or principal components analysis. Then the error matrix Ψ has zero elements outside the diagonal.

The true factor analysis is based on the postulation that not all factors are known. In order to account for this fact the diagonal in the correlation matrix M_R must be reduced i.e. the diagonal elements are

less than 1.0. These diagonal elements are also called "communalities".

Determining Φ and M_A requires a solution for

$$M_R = M_A \Phi M_A^T \quad (3)$$

which is a known problem in mathematics. The model can be reformulated:

$$D_\lambda = M_A^T M_R M_A \quad (4)$$

with $M_A^T = M_A^{-1}$ and D_λ a diagonal matrix. D_λ is called the matrix of eigenvalues and M_A contains the eigenvectors. In the principal components analysis $M_A^T M_A = I$. For more details see Essenwanger (1976).

3. THE COMMUNALITIES. Four different methods have been studied in this investigation. In the first method a principal components analysis (P.C.) is performed and a specific number of factors is accepted. E.g. for a correlation matrix with 8X8 dimension 8 principal component factors are obtained from the mathematical model. We may decide to select the largest 4 factors. This is equivalent to a truncation. The communalities are then recalculated from these 4 accepted factors. This procedure may appear to be somewhat arbitrary and subjective. It must be pointed out, however, that the number of physical factors is unknown. Although the total number of factors in the principal components analysis is determined by the dimension of the matrix M_R the uncertainty of factors with significance in physics is contained in the chosen number of elements in the M_R matrix. A formalistic mathematical solution can be achieved for any dimension of the correlation matrix M_R . However, whether all possible factors in the principal components analysis have significant meaning in physics is not determined by the mathematical solution.

The number of factors is also a subjective choice in the other three methods. Thus the truncation of factors in the principal components analysis is not worse than the assumption of the number of factors in the other three methods.

The other three methods differ how estimators are calculated for the communalities. We assume the number of factors which are

accepted and obtain estimators as follows.

The unweighted least squares method (ULSQ) requires that U is a minimum for

$$U = (1/2) \text{tr} (M_S - M_X)^2 \quad (5)$$

where M_S is the correlation matrix with estimators in the diagonal and tr means the trace.

In the generalized least squares method (GLSQ) G is a minimum for

$$G = (1/2) \text{tr} (I_n - M_S^{-1} M_X)^2 \quad (6)$$

where I_n denotes a diagonal matrix of unity and M_S and M_X are the same as under Eqn (5).

Finally, the maximum likelihood principle (MXLI) is applied to minimize:

$$M = \text{tr} [(M_X^{-1} M_S) - \ln(M_X^{-1} M_S)] - n \quad (7)$$

(See Jöreskog, 1967) where n is the number of variables.

Other methods to substitute estimators for the diagonal in M_R exist (see Essenwanger, 1976) but were not included in the present study; see also Guttman (1956).

4. ROTATIONS. Although the solution of M_A provides characteristic factors which may have meaningful interpretation in physics, it is customary to enhance certain features. This is accomplished by rotation of the coordinate system. This is called attaining simple structure. The ultimate goal is the following:

- (a) At least one zero in each row
- (b) k zeros in each column ($k-1$ for principal components)
- (c) For any pair of factors:

- 1. High loading in one element ~ 1.0

2. Zero in other variables
3. Small loading on both factors for the variable
4. Only a few non-vanishing loading on both.

In order to explain the rotation procedure let us recall that:

$$M_A = M_E D_{\lambda} \quad (8)$$

where M_E is an eigenvector matrix and D_{λ} is a diagonal matrix of eigenvalues λ , with $\lambda_1 = \sqrt{\lambda}$. Two methods of rotations are customary: orthogonal and oblique rotation. In terms of mathematics the orthogonal rotation is achieved by

$$M_{FO} = M_A T_1 \quad (9)$$

where T_1 is a transformation matrix. Oblique rotation requires two transformation procedures because factor pattern and factor structure matrix are not identical as in the orthogonal transformation.

Thus:

$$M_{FP} = M_A T_2^{-1} \quad (\text{factor pattern matrix}) \quad (10a)$$

$$M_{FS} = M_A T_2 \quad (\text{factor structure matrix}) \quad (10b)$$

While the factors are uncorrelated in the solution of Eqn's 4-7 and the orthogonal rotation, the oblique rotation introduces factors which are correlated. Thus M_{FP} represents the regression coefficients in the structure pattern, and M_{FS} the covariances between variables and factors. The factor pattern is:

$$X_i = a_{i1}f_1 + a_{i2}f_2 + \dots + (e_i) \quad (11)$$

where M_{FP} determines the a_{ij} and M_{FS} the f_j terms; e_i is the error.

5. EIGENVALUES, FACTOR LOADS AND COMMUNALITIES. The introduced four methods of estimating the communalities have been applied to

atmospheric data of Stuttgart (Fed. Rep. Germany). The data cover the period * Sept 1946-August 1953. Eight meteorological elements have been selected: ceiling (CEIL), visibility (VIS), wind direction (WD), windspeed (WS), temperature (TEMP), dewpoint (DEWP), relative humidity (REHU) and pressure (PRES). Visibility was utilized in linear scale and as transformed variate in logarithmic scale. The wind velocity was also converted to zonal (U) and meridional (V) components. These differences in the element selections will be discussed later.

Data as exhibited in Tables 1 and 2 were chosen as a typical example for disclosing the diversity caused by different methods of estimating the communalities. Table 1 displays the eigenvalues for data from Stuttgart (linear visibility, zonal and meridional wind components). We learn from perusal of Table 1 that the individual eigenvalues fluctuate and depend on the chosen method. The dissimilarity is even found in the sums of these eigenvalues. However, rotation of the coordinate systems (orthogonal and oblique) has no effect on the sum, as expected. The numerical values differ only by rounding.

The differences between the individual methods for the sum of eigenvalues can be traced to the sum of communalities (Table 2). As confirmed by the observed data the sum of eigenvalues must be identical with the sum of the communalities save rounding. In the principal components analysis this sum is identical with the number of elements if the number of factors is not truncated.

We also notice in Table 1 that the truncated principal components analysis shows the highest approximation (82%) of the total variance for the chosen number of factors, in our case four.

*Footnote: We experienced difficulty with the magnetic tape record after 7 years of data. The difficulty could not be resolved for inclusion into this manuscript. Only Table 3 was available for 10 years.

While Table 1 exhibits fluctuations of the sum of eigenvalues from 6.567 to 5.040 these variations are not necessarily repeated for other data sets. E.g. Table 3 has been compiled for 10 years of data for Stuttgart in January, substituting visibility in its transformed logarithmic scale, and zonal and meridional components of wind have been replaced by speed and direction (see Essenwanger, 1964). We learn that the sum of the eigenvalues for the three methods ULSQ, GLSQ, and MXLI differ very little, although the individual eigenvectors show dispersion. Again, the truncated principal components analysis renders the highest approximation of the variance (about 81%).

6. FACTOR LOADS, STRUCTURE MATRIX AND FACTOR PATTERN. Tables 4A-D provide detailed information about the factors. Four sections are shown in each Table 4A-D. The first section provides the unrotated factor loads for the solution with communalities. E.g. in the case of the principal components method (Table 4A) these are the first 4 eigenvectors of a correlation matrix with unity in the diagonal matrix. The numerical values in these four columns represent the affinity with the elements and can be interpreted as a (linear) correlation coefficient.

The first factor (Table 4A) which represents 39% of the variance (i.e. $3.14/8.00$) discloses high association with temperature, dewpoint, zonal (U) and meridional (V) wind component and visibility, in that order of magnitude. The second factor with about 21% of the variance is again a mixture, relative humidity, visibility, dewpoint and ceiling. In the third factor the pressure stands out while the fourth factor is again a mixture whereby all elements are contributing except the relative humidity ($-.07$ means almost zero).

The unrotated factor load is a valid solution. It was pointed out previously that a rotation of the coordinates will enhance the association between individual factor and element. This simplification process was described in section four. The sum of the eigenvalues remains constant in this transformation.

Inspection of the section for orthogonal rotation in Table 4A reveals that now the first factor principally is related with the temperature elements, i.e. temperature and dewpoint. The second factor comprises the moisture elements (relative humidity, visibility and ceiling). The third factor contains the pressure, and the fourth factor the wind. This may be expected by some readers and may be a trivial answer. It should be stressed, however, that the mathematical formalism could have led to

a different answer and combination of elements. The separation into these four factors is logical on account of the physics background. This may give the impression that the grouping into these 4 factors is trivial. In turn, the mathematical formalism has led in this case to an answer which has an interpretation in terms of physics. However, beyond the expected factors we gain information about the weights of the factors. This weight is not readily available by expectation alone.

The lower part of Table 4A lists the result for an oblique rotation. While the structure matrix contains the covariances (which are equivalent to the correlation coefficient); the factor pattern expresses the regression coefficients. In the oblique rotation the factors are intercorrelated (see Table 5). They are not correlated with each other for the unrotated or the orthogonal solution. We learn from the structure matrix of Table 4A that the factors have not essentially changed from the orthogonal rotation case. Therefore, the intercorrelation (between factors) is very low (Table 5).

The results for the other methods (ULSQ, GLSQ, MXLI) are similar with minor changes except that the weights are different for the individual factors. In Table 4B we notice that the ceiling shows only very low influence in any of the factors. This result is repeated in Table 4C. While in the previous methods the pressure is one factor, it shows virtually no contribution in the GLSQ method. It reappears as a factor in Table 4D, MXLI method. Another difference between Tables 4A, B and Tables 4C, D is the influence of the windspeed. In Table 4A the factor with the two wind components indicates equal correlation of the wind components. In Table 4B a small preference of the meridional component is already visible. In Tables 4C, D, however, the meridional wind component appears to be more dominant than the zonal influence in the wind factor.

One further peculiarity must be mentioned. In the unrotated and orthogonally rotated case the sum of the eigenvalues SU_{λ} AND SO_{λ} , respectively, is equal to the sum of the squares of the factor components.

$$SU_{\lambda} = \sum_i^n f_u^2 \quad (12a)$$

$$\text{or } SO_{\lambda} = \sum_i^n f_o^2 \quad (12b)$$

where f_u^2 and f_o^2 denote the numerical value in the respective

factor column and n designates the number of elements. In the oblique case we find

$$SOB_{\lambda} = \sum_i^n f_s f_F \quad (13)$$

where f_s is the column value in the structure matrix and f_F the corresponding column value in the factor pattern. Although the sum of SOB_{λ} for the 4 factors renders the same numerical value as the unrotated or orthogonally rotated case the individual items SOB_{λ} can be positive or negative in the maximum likelihood method (Table 6). The exhibited case in Table 6 is not an isolated case or error as the first impression may be. As can be seen from Table 7A a negative term appears also in a combination of elements Ln VIS, WD, WS. In July (Table 7B) this peculiarity did not show, and it almost rules out that it is an error in the computer program. Thus the maximum likelihood method, at least in our "canned computer program", appears to be very sensitive to changes of the correlations in the input matrix.

7. FACTOR ANALYSIS. The detailed information on unrotated and rotated factors is listed in Tables 4A-D for one version of a set of elements. These detailed tabulations are somewhat difficult to read. In order to enhance the significant features of the factors, two changes were introduced for Tables 7A and B. First, all correlations $r \leq 0.4$ were omitted except the maximum correlation in one line which could be smaller than 0.4. Secondly the sign was omitted because the sign plays only a role in formulating eqn (11) and performing calculations with it. The magnitude is sufficient for evaluation of the factors.

In Table 7A, B eight atmospheric elements are shown. For these eight elements visibility was used in its linear scale and with a transformed (logarithmic) scale. In the top part of Tables 7A, B the wind appears as speed and direction while in the center and lower section the zonal and meridional components have been utilized. These modifications lead to three different versions of factor analysis for the same elements. Only the solutions with orthogonal and oblique rotation are included in Tables 7A, B.

Table 7A exhibits the condition for January. The significant features do not vary essentially between the three versions. The only exception is the contribution by ceiling of clouds which renders a significant factor for the ULSQ method (top and center) but is not a special factor at the bottom section where it is replaced by the pressure. The differences between individual methods (PC, ULSQ, GLSQ,

and MXLI) were mostly described in the previous section 6 and will not be repeated here.

Table 7B provides the factor analysis for July at Stuttgart for the same seven-year period of record at Stuttgart. Again, it can be noticed that the oblique rotation is not significantly different from the factors provided by orthogonal rotation. Other data, not included here, follow the same trend that orthogonal and oblique rotation do not differ significantly. This fact may imply that orthogonal rotation may be sufficient for factor analysis of atmospheric elements. Although the characteristic of factors shows a similar pattern in July as given for January, some difference exist. Besides the mentioned difference in the contribution by the ceiling a major change has occurred in the association of elements. Relative humidity and visibility are now associated with temperature in three of the four methods for all three versions. This first factor proves to be the dominant influence but not by much.

The primary purpose of this study was not the illustration of the changes throughout the year but the exhibition of the differences in the utilization of the individual methods. Although variations exist, a close perusal reveals that physical characteristics of the system do not differ too much in the individual methods.

8. CONCLUSION AND SUMMARY. The present study illustrates that the estimation approach for the communalities by different methods (eqn 5-7) leads to different factors. They are more uniform, however, after rotation of the factors. This confirms that the basic problem in factor analysis has not been resolved as of today, namely the derivation of suitable estimators for the communalities (see Cattell, 1965 or Guttman, 1956). As the study proves, however, the physical features after rotation of the factors show major agreement, although differences in details and in the sequence of importance of factors can be found.

This study revealed that for atmospheric elements the factors derived by oblique rotation do not differ significantly from factors procured by orthogonal rotation. This may imply that the elaborate mathematical procedure for oblique rotation could be saved in favor of the simpler and less costly orthogonal rotation.

The factors appearing in the January data are related to four simple combinations, temperature, wind, moisture and pressure. This simple division is not repeated in the July data. However, the resulting factors from the analysis procedure do not give unreasonable combinations in terms of physics. E.g. the combination of temperature with visibility and relative humidity may have some explanation in terms of relationship between reduced radiation during high relative humidity and low visibility and vice versa. Also the combination of a wind component with

temperature terms may indicate a reflection of the circulation of air either in the macro- or meso-scale. Other detailed features in the patterns of factors may be reserved for a further study.

Finally, no specific recommendation as to the "best suitable method" for estimating the communalities can be made at the present time.

9. REFERENCES. Buell, C. E., 1971. Integral Equations Representation for Factor Analysis. J. Atmosph. Sci., 28, 1502-1505.

Cattell, R. B., 1952. Factor Analysis. Harper, New York, pp. 462.

Cattell, R. B., 1965. The Configurative Method for Surer Identification of Personality Dimensions, Notably in Child Study Psych. Rep, 16, 269-270.

Christensen, W. I, Jr. and R. A. Bryson, 1966. An investigation of Component Analysis and Weather Classification. Mon. Weather Rev, 94, 697-709.

Essenwanger, O. M., 1964. The Cumulative Distribution for Wind Direction Frequencies. Meteorol. Rundsch. 17, 131-134.

Essenwanger, O. M., 1976. Applied Statistics in Atmospheric Science Elsevier, Amsterdam, pp. 412.

Guttman, L., 1956 "Best Possible" Systematic Estimates of Communalities Psychometrika, 21, 273-285.

Jöreskog, K. G. (1967). Some Contributions to Maximum Likelihood Factor Analysis. Psychometrika, 32, 443-482.

Kutzbach, J. E., 1967. Empirical Eigenvectors of Sea-Level Pressure Surface Temperature and Precipitation Complexes Over North America, J. Appl. Meteorol., 6, 791-802.

Spearman, C., 1904. General Intelligence Objectively determined and measured. Am. J. Psychol, 15, 201-293.

Spearman, C., 1927. The Abilities of Man. Mac Millan, London, pp. 87.

ACKNOWLEDGEMENT: The author's thanks go to Dr. Dorothy A. Stewart for her critical review of the manuscript. Mrs. Alexa Mims and Mr. Roger Betts deserve the credit for the preparation of the computer programs to obtain the data. Last, not least, Mrs. Gloria McCrary must be thanked for her patience and diligence during the process of establishing the manuscript from the first draft to the final version.

TABLE 1. COMPARISON OF EIGENVALUES, FACTOR LOADS
(STUTTGART, JANUARY)

(1) Unrotated Factor Loads

	PC	ULSQ	GLSQ	MXLI
λ_1	3.136	2.929	2.811	2.303
λ_2	1.695	1.385	1.590	1.462
λ_3	1.016	0.924	0.636	1.328
λ_4	0.720	0.432	0.003	.789
$\Sigma \lambda$	6.567	5.670	5.040	5.882

(2) Orthogonal Factor Load

λ_1	2.150	2.157	2.252	2.196
λ_2	1.611	1.152	1.257	1.076
λ_3	1.200	1.080	1.528	1.272
λ_4	1.601	1.273	0.003	1.337
$\Sigma \lambda$	6.562	5.662	5.040	5.881

(3) Oblique Structure Matrix

λ_1	2.128	2.102	2.192	1.189
λ_2	1.613	1.170	1.262	2.238
λ_3	1.203	1.081	1.576	1.460
λ_4	1.622	1.311	0.011	.994
$\Sigma \lambda$	6.566	5.664	5.041	5.881

TABLE 2. COMMUNALITIES

(STUTTGART, JANUARY)

	PC	ULSQ	GLSQ	MXLI
CEIL	.697	.234	.159	.200
VISIB	.758	.504	.399	.428
U	.729	.507	.424	.477
V	.811	.714	1.000	.781
TEMP	.947	1.002	1.000	.996
DEWP	.988	1.007	1.000	1.000
REHU	.749	.693	1.000	.999
PRES	.887	1.002	.058	1.000
Σx^2	6.566	5.663	5.040	5.881

TABLE 3. EIGENVALUES AND COMMUNALITIES

STUTTGART, JANUARY, 1946-1956, Ln Vis, WDD, WSP
(A) EIGENVALUES (ORTHO. FACT. LOAD)

	PC	ULSQ	GLSQ	MXLI
λ_1	2.207	1.863	2.042	1.868
λ_2	2.053	1.532	1.310	1.525
λ_3	1.254	1.185	1.230	1.188
λ_4	1.004	1.062	1.018	1.063
$\Sigma \lambda$	6.518	5.642	5.600	5.642

(B) COMMUNALITIES

	PC	ULSQ	GLSQ	MXLI
	.802	1.000	.146	1.000
	.740	.532	.441	.531
	.630	.498	1.000	.501
	.712	.592	1.000	.591
	.941	.990	1.000	.990
	.996	1.000	.995	1.000
	.705	1.000	1.000	1.000
	.991	.031	.018	.031
Σx^2	6.517	5.643	5.600	5.644

TABLE 4A. FACTOR LOADS, STRUCTURE MATRIX AND FACTOR PATTERN
(STUTTGART, JANUARY)
PRINCIPAL COMPONENTS

	UNROTATED				ORTHOG. ROT.			
	PC	ULSQ	GLSQ	MXLI	PC	ULSQ	GLSQ	MXLI
CEIL	.44	-.46	.48	.24	.19	-.49	.54	.37
VIS	-.59	-.58	.05	.26	-.35	-.73	-.18	-.26
U	-.76	-.04	.10	-.38	-.40	-.05	-.05	-.75
V	.67	.38	.03	.47	.12	.29	.10	.84
TEMP	-.87	.18	.27	.28	-.92	-.12	-.11	-.26
DEWP	-.80	.47	.26	.23	-.95	.18	-.09	-.20
REHU	.10	.86	.00	-.07	-.21	.82	.02	.16
PRES	.39	.09	.80	-.30	.08	.16	.92	.03
$\Sigma \times^2$	3.14	1.70	1.02	.72	2.15	1.61	1.20	1.60

OBLIQUE ROTATION

	STRUCTURE MATRIX				FACTOR PATTERN			
CEIL	.28	-.44	.58	.39	.10	-.50	.52	.36
VIS	-.37	-.75	-.21	-.38	-.32	-.73	-.13	-.16
U	-.49	-.11	-.13	-.81	-.29	-.00	.01	-.73
V	.22	.36	.17	.87	-.02	.23	.05	.83
TEMP	-.95	-.14	-.18	-.42	-.91	-.14	-.04	-.14
DEWP	-.98	.16	-.17	-.34	-.95	-.16	-.04	-.09
REHU	-.21	.83	.01	.19	-.24	.81	.01	.13
PRES	.13	.19	.92	.11	.03	.19	.93	-.06

Structure Matrix = Covariance

Factor Pattern = Regression Coefficients

TABLE 4B. FACTOR LOADS, STRUCTURE MATRIX AND FACTOR PATTERN
STUTTGART, JANUARY
UNWEIGHTED LEAST SQUARE

	PC	UNROTATED				PC	ORTHOG. ROT.		
		ULSQ	GLSQ	MXLI			ULSQ	GLSQ	MXLI
CEIL	.35	-.23	.21	-.12		.29	-.25	.23	-.19
VIS	.51	-.47	.12	-.11		-.25	-.54	-.14	.36
U	-.66	-.73	.07	.26		-.39	-.05	-.10	.58
V	.61	.40	-.09	-.42		.15	.23	.10	-.79
TEMP	-.92	.19	.19	-.28		-.95	-.17	-.07	.25
DEWP	-.85	.51	-.09	-.11		-.95	.21	-.07	.22
REHU	-.07	.76	-.21	.25		-.15	.81	.00	.12
PRES	.40	.26	.87	.11		.07	-.01	.99	-.11
Σx^2	2.93	1.38	0.92	0.43		2.16	1.15	1.08	1.27

	STRUCTURE MATRIX					FACTOR PATTERN			
CEIL	.32	-.23	.26	-.23		.23	-.27	.22	-.18
VIS	-.29	-.58	-.18	.46		-.22	-.53	-.10	.26
U	-.46	-.11	-.16	.66		-.28	.00	-.04	.56
V	.24	.33	.16	-.83		-.01	.16	.04	-.79
TEMP	-.97	-.16	-.15	.46		-.94	-.18	-.03	.11
DEWP	-.98	.22	-.14	.40		-.92	.20	-.03	.14
REHU	-.15	.82	.02	-.16		-.12	.81	.00	-.03
PRES	.12	.11	.99	.21		.01	.06	.99	-.04

Structure Matrix = Covariance
Factor Pattern = Regression Coefficients

TABLE 4C. FACTOR LOADS, STRUCTURE MATRIX AND FACTOR PATTERN
STUTT GART, JANUARY

	GENERAL				LEAST SQUARES			
	UNROTATED				ORTHO. ROT			
	PC	ULSQ	GLSQ	MXLI	PC	ULSQ	GLSQ	MXLI
CEIL	-.36	.14	.10	.03	-.33	.16	.15	.02
VIS	.39	.49	.08	.01	.24	.44	-.38	.01
U	.60	.19	-.17	-.003	.42	.06	-.49	-.003
V	-.54	-.61	.57	.000	-.11	-.15	.98	.01
TEMP	.96	.03	.26	-.03	.94	.21	-.25	-.02
DEWP	.95	-.29	.10	.03	.97	-.14	-.19	.04
REHU	.09	-.90	-.42	-.01	.21	-.97	.15	.005
PRES	-.20	-.12	.06	-.01	-.13	-.06	.20	-.004
Σx^2	2.81	1.59	0.64	0.003	2.25	1.26	1.53	.003

	STRUCTURE MATRIX				FACTOR PATTERN			
	PC	ULSQ	GLSQ	MXLI	PC	ULSQ	GLSQ	MXLI
CEIL	-.34	.17	.18	-.03	-.30	.18	.14	.03
VIS	.26	.46	-.44	-.19	.22	.43	-.33	.007
U	.46	.08	-.55	-.10	.36	.03	-.47	-.008
V	-.18	-.22	.99	.31	.02	-.07	.98	.01
TEMP	.96	.18	-.39	-.02	.94	.20	-.17	-.03
DEWP	.98	-.18	-.31	.15	.94	-.13	-.15	.04
REHU	.21	-.99	.19	.35	.16	-.96	.08	.004
PRES	-.14	-.07	.22	.05	-.10	-.05	.19	-.003

Structure Matrix = Covariances

Factor Pattern = Regression Coefficient

TABLE 4D. FACTOR LOADS, STRUCTURE MATRIX, FACTOR PATTERN
STUTTGART, JANUARY

MAXIMUM LIKELIHOOD

	UNROTATED				ORTHO. ROT			
	PC	ULSQ	GLSQ	MXLI	PC	ULSQ	GLSQ	MXLI
CEIL	.42	-.09	.11	-.08	.31	.23	.17	-.14
VIS	-.31	.00	.52	.24	-.24	-.15	.43	.40
U	-.46	.23	.20	.42	-.41	-.10	.04	.55
V	.32	-.05	-.35	-.74	.14	.09	-.17	-.85
TEMP	-.73	.58	.35	-.00	-.95	-.06	.17	.25
DEWP	-.76	.65	.00	.00	-.96	-.06	-.18	.19
REHU	-.16	.29	-.94	.00	-.16	-.01	-.98	-.14
PRES	.76	.65	.00	.00	.08	.99	-.07	-.12
$\Sigma \times^2$	2.30	1.33	1.46	.79	2.20	1.08	1.27	1.34

	STRUCTURE MATRIX				FACTOR PATTERN			
	PC	ULSQ	GLSQ	MXLI	PC	ULSQ	GLSQ	MXLI
CEIL	.25	-.36	-.11	-.10	.22	-.43	.16	-.08
VIS	-.24	.23	.56	.39	-.22	-.14	.55	.25
U	-.18	.49	.43	.53	-.14	.35	.11	.44
V	.21	-.27	-.43	-.83	.19	-.02	-.20	-.78
TEMP	-.15	.93	.81	.23	-.08	.68	.42	.001
DEWP	-.11	1.00	.56	.16	-.04	.99	.00	.00
REHU	.08	.31	-.61	-.18	.10	.96	-.01	-.01
PRES	1.0	-.11	-.06	-.02	.99	-.04	-.00	.00

Structure Matrix = Covariances

Factor Pattern = Regression Coefficients

TABLE 5. INTERCORRELATION BETWEEN FACTORS
(OBLIQUE ROTATION)

A) Principal Components Analysis

1.0	-.02	.13	.28
-.02	1.0	.02	.15
.13	.02	1.0	.15
.28	.15	.15	1.0

B) Unweighted Least Squares

1.0	-.05	.10	-.33
-.05	1.0	.04	-.21
.10	.04	1.0	-.16
-.33	-.21	-.16	1.0

C) General Least Squares

1.0	-.07	-.20	.12
-.07	1.0	-.15	-.32
-.20	-.15	1.0	.27
.12	-.32	.27	1.0

D) Maximum Likelihood

1.0	-.08	-.04	-.01
-.08	1.0	.56	.16
-.04	.56	1.0	.28
-.01	.16	.28	1.0

TABLE 6. VARIANCE COMPONENTS FOR THE MAXIMUM LIKELIHOOD
METHOD (JANUARY, STUTTGART, LN VIS, U, V)

	UNROT	ORTH. ROT	OBLIQUE ROT.
x ₁	2.116	1.102	1.469
x ₂	1.374	1.255	3.934
x ₃	1.448	2.003	12.392
x ₄	0.830	1.407	-12.028
Σx	5.768	5.767	5.767

TABLE 7A, FACTORS FOR JANUARY, STUTT GART (UNIT: PERCENT)

Ortho rot					Oblique rot				
	P.C.	ULSQ	GLSQ	MXLI	P.C.	ULSQ	GLSQ	MXLI	
CEIL									
Ln VIS	61	98	35	98	65	99	34	99	
WD	82	55	46	61	83	44	44	49	
WS	94	33	10		94	62	23	18	
TEMP	43 51	69	95	63	44 54	73	99 66	33	
DEWP	93	96	97	96	95	98	98	95 99	
REHU	97	96	99	96	97	98	98	99 93	
PRES	83	95	96	94	83	99	67	99	
	90	27	29		89	39	20	25	
						29			
λ	2.26 1.24	2.04 1.10	2.24 1.22	1.10 2.03	2.25 1.25	2.04 1.07	2.11 1.18	1.13 4.68	
	1.83 1.08	1.13 1.06	1.25 0.02	1.09 1.09	1.82 1.09	1.15 1.07	1.15 0.29	2.89 -3.39	
CEIL	49 50 43	97	33	97	46 54 44	99	33	99	
Ln VIS	77	47	47	43	78	53	45 45 57	50	
U	41	59	43	49	80	65	54	50 51	
V	82	78	98	80	86	81	100 52	35	
TEMP	93	94	95	94	95	97	49	95 100	
DEWP	96	94	97	94	98	97	97	99 93	
REHU	82	91	97	98	83	93	62	100	
PRES	93	24	19	23	94	26	21	26	
λ	2.10 1.17	2.00 1.09	2.24 1.50	1.10 2.00	2.06 1.17	1.99 1.07	2.12 1.53	1.47 12.39	
	1.69 1.60	1.20 1.37	1.28 0.02	1.26 1.41	1.68 1.65	1.21 1.38	1.16 0.24	3.93 -12.02	
CEIL	49 54	29	33	31	44 58	32	34	36	
VIS	73	54	44	43 40	75	58	46 44	56	
U	40	58	42	41	55	46	46 55	49 43 53	
V	84	79	98	85	88	83	99	43 84	
TEMP	92	95	94	95	42	97	96	93 81	
DEWP	95	95	97	96	98	98	40	100 56	
REHU	82	81	97	98	83	82	35	61	
PRES	92	99	20	99	92	99	22	100	
λ	2.15 1.20	2.16 1.08	2.25 1.53	2.20 1.27	2.13 1.20	2.10 1.08	2.19 1.58	2.20 1.27	
	1.61 1.60	1.15 1.27	1.26 .003	1.08 1.34	1.61 1.62	1.17 1.31	1.26 .01	1.08 1.34	

$|r| \geq 0.40$ or $\{max\}$

TABLE 7B, FACTORS FOR JULY, STUTTGART, (UNIT: PERCENT)

ORTHO ROT					OBLIQUE ROT				
	P.C.		ULSQ		GLSQ		MXLI		
CEIL.		69		33	28		95		
LnVis	70		42		42		41		
WD		46	26			21		28	
WS		83		98	12			74	
TEMP	64	70	80	58	69	68	24	56	80
DEWP		93		99		98		99	
REHU	91		98		98		99		
PRES		94		99		99		13	
λ	2.01	1.57	1.95	1.10	1.78	1.02	1.37	.95	
		1.33	1.38	1.00		1.49	1.11	1.88	.73
CEIL		66		42	27		43		
LnVis	70			97	42		64		
U		80		70		24		71	
V		81		78	13			77	
TEMP	69	67	82	55	69	68	23	54	80
DEWP		96		99		98		99	
REHU	91		94		99		93		
PRES		99		13		99		12	
λ	2.00	1.46	1.75	1.36	1.79	1.02	1.36	1.33	
		1.79	1.32	1.00		1.50	1.01	1.71	.54
CEIL		67		51	27		44		
VIS	67		40		37		43		
U		80		89		24		66	49
V		79		56	13			75	
TEMP	68	67	79	57	70	70	20	76	
DEWP		96		99		98		99	
REHU	91		97		99		96		
PRES		99		15		99		13	
λ	1.95	1.46	1.82	1.36	1.77	1.02	1.39	1.23	
		1.82	.99	.81		1.50	.06	1.80	.36

|r| ≥ 0.40 or |max|

Small Composite Designs

Norman R. Draper

Statistics Department
University of Wisconsin
Madison, WI 53706

Small second-order composite designs were suggested by Hartley (1959). Westlake (1965) provided even smaller designs for $k = 5, 7$, and 9 factors, for which intricate construction methods were needed. Here, simple designs formed using Plackett and Burman (1946) designs are offered for $k = 5, 7$, and 9 . Designs with one run fewer than Westlake's for $k = 5$ and 7 and three fewer for $k = 9$ are feasible by deleting repeat points that occur in some of the designs.

KEY WORDS: Center points; Composite designs; Factorial designs; Plackett and Burman designs; Response surfaces.

1. INTRODUCTION

Suppose we are going to examine k predictor variables, coded to x_1, x_2, \dots, x_k , to determine their effects on a response variable y subject to random error. We might first wish to perform a first-order design to fit the model $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$. If no progress appeared possible (for example, via steepest ascent), we might then wish to add a few runs to enable the more comprehensive second-order model,

$$y = \beta_0 + \sum \beta_i x_i + \sum_{i \geq j} \beta_{ij} x_i x_j + \varepsilon, \quad (1)$$

to be examined, where all summations are taken over $i, j = 1, 2, \dots, k$. Many possible second-order sequential designs may be used to obtain the data for such a fitting. The specific choice of design would depend on the relative importance to the experimenter of various design features (for example, see Box and Draper 1975, p. 347). One extremely useful type of sequential second-order design is the *composite* design. As initially suggested by Box and Wilson (1951) and followed up by Box and Hunter (1957), it consists of a 2^k factorial or a 2^{k-q} fractional factorial portion, with runs selected from the 2^k runs $(x_1, x_2, \dots, x_k) = (\pm 1, \pm 1, \dots, \pm 1)$, of resolution V or higher (for example, see Box and Hunter 1961 or Box, Hunter, and Hunter 1978), plus a set of $2k$ axial points at distances α from the origin, plus n_0 center points. In general, the 2^{k-q} portion or *cube* may be repeated c times, and the axial points or *star* may be repeated s times. The values of α , n_0 , c , and s are to be selected.

Suppose, of the various design criteria, we decide to emphasize having only a small number of runs. Such a course of action might be appropriate if runs were expensive, difficult, or time-consuming, or if a complicated computer model were to be approximated locally by a second-order surface. Of course there must

be at least $\frac{1}{2}(k+1)(k+2)$ points in the design, this being the number of coefficients to estimate in (1). Hartley (1959) pointed out that the cube portion of the composite design need not be of resolution V. It could, in fact, be of resolution as low as III, provided that two-factor interactions were not aliased with two-factor interactions. (Two-factor interactions could be aliased with main effects, because the star portion provides additional information on the main effects.) This idea permitted much smaller cubes to be used. Westlake (1965) took this idea further by finding even smaller cubes for the $k = 5, 7$, and 9 cases. Table 1 shows the numbers of points in the various designs suggested, for $2 \leq k \leq 9$.

Westlake (1965) provided (in an appendix) three examples of 22-run designs for $k = 5$, one example of a 40-run design for $k = 7$, and one example of a 62-run design for $k = 9$. He noted that for $k = 7$ or 9 , "systematic generation of all possible designs ... appears to be almost out of the question" (p. 332).

Table 1. *Points Needed by Some Small Composite Designs*

Factors, k	2	3	4	5	6	7	8	9
Coefficients								
$\frac{1}{2}(k+1)(k+2)$	6	10	15	21	28	36	45	55
Points in Box-Hunter (1957) designs	8	14	24	26	44	78	80	146
Hartley's number of points	6	10	16	26	28	46	48	82
Westlake's number of points	—	—	—	22	—	40	—	62

2. CONSTRUCTING SMALL COMPOSITE DESIGNS

Can Westlake's small numbers of runs for the $k = 5$, 7, and 9 cases be beaten? The surprising answer is yes. Moreover, for $k = 5$ and 9 it is possible to *equal* the number of runs in a simple manner, and for $k = 7$, simple designs are available with only 42 runs, two more than Westlake's 40. The overall advantage of these suggested designs is that none of the ingenuity shown by Westlake (1965) is needed, thanks to Plackett and Burman (1946), and yet an apparently large selection of possibilities is immediately available. (As we shall see later, the selection is not as large as first appears!)

The basic method can be simply stated: (a) Use, for the cube portion of the design, k columns of a Plackett and Burman (1946) design. (b) Where repeat runs exist, remove one of each duplicate pair to reduce the number of runs.

Let (1) be written in the matrix form $y = X\beta + \epsilon$. If $(X'X)^{-1}$ exists, we have a valid second-order response-surface design that will estimate all of the parameters in (1). To avoid the possibility of actual or near singularity merely due to choice of α , I initially followed Westlake (1965) by selecting the star with unit axial distance, namely with points $(\pm 1, 0, \dots, 0)$, $(0, \pm 1, \dots, 0)$, \dots , $(0, 0, \dots, \pm 1)$. In practice, this value of α may be varied, since its value does not affect the singularity or nonsingularity of the design, apart from the following feature: When $\alpha \neq k^{1/2}$, the design has two spheres of points with radii $k^{1/2}$ and α , so center points are not needed (see Box and Hunter 1957, p. 217). If the choice $\alpha = k^{1/2}$ were made, however, center points would be essential to avoid design singularity. In later computations reported here, I used the values $\alpha = 2$ (for $k = 5$), $\alpha = 8^{1/2} = 2.828427$ ($k = 7$), and $\alpha = 2^{7/4} = 3.363586$ ($k = 9$). These were suggested by a referee, because they are the values that provide rotatable designs if a 2^{k-1} design is used with a star of axial distance α for $k = 5$ and 7, and if a 2^{k-2} design is used similarly for $k = 9$.

3. CASE $k = 5$

There are 21 coefficients to estimate, and there are 10 axial points. The difference of 11 is thus the minimum possible number of cube points required. An obvious choice is to use five (of the 11) columns of a 12-run Plackett and Burman (1946) design. There are $\binom{11}{5}$ that is, 462 possible choices, all of which produce nonsingular designs. These require 22 runs, the same number as Westlake's. A detailed examination of the cube portions for the designs shows that there are two basic types; standardized versions of these appear in Table 2.

Table 2. Two Essentially Different Choices of Five Columns From a 12-Run Plackett and Burman Design: (a) With a Pair of Repeat Runs; (b) With a Mirror-Image Pair of Runs

a					b				
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	+	+	+	+	+
-	-	+	+	+	-	-	+	+	+
-	+	-	+	+	-	+	+	-	+
-	+	+	-	+	+	-	-	+	+
-	+	+	+	-	+	+	-	+	-
+	-	-	+	+	+	+	+	-	-
+	-	+	-	+	+	-	+	-	-
+	-	+	+	-	+	-	-	-	+
+	+	-	-	+	-	+	-	+	-
+	+	-	+	-	-	+	-	-	+
+	+	+	-	-	-	-	+	+	-

NOTE: All other choices are equivalent to one of these, subject to changes in signs throughout one or more columns, renaming of variables, and reordering of runs.

4. CASE $k = 7$

There are 36 coefficients to estimate, and there are 14 axial points. Thus a minimum of 22 cube points is needed. First an attempt was made to form designs using seven (of the 23) columns of the 24-run Plackett and Burman design. Tries with columns (1-7), (1, 2, 4, 5, 8, 9, 10), (3-5, 7-10), and (1, 3, 4, 7-10) all produced singular $X'X$ matrices. There are, in all, 245,157 possible column choices, and it is conjectured that all will fail.

A second attempt used seven (of the 27) columns of the 28-run Plackett and Burman design. More than 20 tries all produced nonsingular designs with no failures, and it is conjectured that all of the 888,030 choices of seven columns from 27 will do the same. These designs have 42 runs, a modest two more than Westlake's 40, but reduced designs with fewer runs are also possible.

Features we have already noted in the $k = 5$ case also arise here. Many of the possible column choices provide identical or essentially identical sets of points; some choices provide repeat runs and some provide mirror-image runs. A new feature for $k = 7$ is that some sets of columns provide both repeats and mirror images, and some neither!

How many distinct designs are there? Based on the number of different $|X'X|$ matrices found in a trial-and-error selection of designs, there are at least 15.

5. CASE $k = 9$

There are 55 coefficients to estimate, and there are 18 axial points. Thus a minimum of 37 cube points is needed. One possibility is to use nine (of the 39) columns of the 40-run Plackett and Burman design. Tries with columns (1-9) and (2-9, 39) failed, producing a singular $X'X$ matrix. It is conjectured that all 211,915,312 possible choices will fail similarly. Parallel to this, I note Westlake's (1965) remark that, for a $3/16$ fraction of a 2^7 , "while one apparently valid defining relation exists, it is impossible to pick three $1/16$ replicates so as to give a non-singular $X'X$ matrix" (p. 329).

A second attempt used nine (of the 43) columns of the 44-run Plackett and Burman design. More than 20 tries all produced nonsingular 62-run designs, the same number of runs as Westlake's. There were no failures, and it is conjectured that all 563,921,995 column choices will produce nonsingular designs.

Features similar to the $k = 7$ case again arise. Designs certainly exist with up to three pairs of repeats and up to two pairs of mirror-image runs.

ACKNOWLEDGMENTS

Partial sponsorship was provided by the University of Wisconsin Graduate School through funds provided by the Wisconsin Alumni Research Foundation, and by the U.S. Army Research Office under Contract DAAG29-80-C-0041.

REFERENCES

- BOX, G. E. P., and DRAPER, N. R. (1975), "Robust Designs," *Biometrika*, 62, 347-352.
- BOX, G. E. P., and HUNTER, J. S. (1957), "Multi-Factor Experimental Designs for Exploring Response Surfaces," *Annals of Mathematical Statistics*, 28, 195-241.
- (1961), "The 2^{k-p} Fractional Factorial Designs" (Parts I and II), *Technometrics*, 3, 311-351 and 449-458.
- BOX, G. E. P., HUNTER, W. G., and HUNTER, J. S. (1978), *Statistics for Experimenters: An Introduction to Design, Data Analysis and Model Building*, New York: John Wiley.
- BOX, G. E. P., and WILSON, K. B. (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society, Ser. B*, 13, 1-45.
- HARTLEY, H. O. (1959), "Smallest Composite Designs for Quadratic Response Surfaces," *Biometrics*, 15, 611-624.
- PLACKETT, R. L., and BURMAN, J. P. (1946), "The Design of Optimum Multi-Factorial Experiments," *Biometrika*, 33, 305-325.
- WESTLAKE, W. J. (1965), "Composite Designs Based on Irregular Fractions of Factorials," *Biometrics*, 21, 324-336.

This note is reduced and adapted from the author's paper in Technometrics, 27, May 1985, pp. 173-180. Please refer to that paper for omitted details; reprints are available from the author. I am grateful to the copyright holder, American Statistical Association, for permission to reproduce in the present form.

CONSIDERATIONS IN SMALL SAMPLE QUANTAL RESPONSE TESTING

Barry A. Bodt
U.S. Army Ballistic Research Laboratory
ATTN: SLCBR-SE-D
Aberdeen Proving Ground, MD

Henry B. Tingey
University of Delaware
Newark, DE

ABSTRACT

In the army sensitivity testing environment it is often desired to estimate V_{50} , the velocity at which 1/2 of a given projectile population would penetrate a given plate of armor. Excessive cost of experimental units usually necessitates the use of very small samples - often less than 15. Several studies have been done to examine the performance of some of the available design and estimation techniques under restrictive sample sizes. Discussed will be some extensions of those studies with emphasis on additional practical environment considerations such as nonnormal response functions, stimulus noise, estimate existence, and initial design point selection.

INTRODUCTION

In the army quantal response testing environment, excessive cost of experimental units usually necessitates the use of small samples. Several small sample studies have been done to examine the performance of some of the available design and estimation techniques. This paper discusses extensions of those studies including additional practical environment considerations such as estimate existence, nonnormal response functions, and stimulus noise.

The quantal response testing environment is one in which there are only two possible outcomes for each experimental unit. For example, if a projectile were fired against a plate of armor one could observe a penetration (response) or a nonpenetration. Continuing with this example, suppose an experimenter wishes to assess the performance of a particular projectile. One way to characterize performance is to consider the probability of a projectile perforating the armor at various velocities. Thus, assessing the performance of a projectile in this manner amounts to establishing some appropriate probability distribution.

Assume that associated with every projectile is a critical velocity above which the projectile would penetrate the armor and below which it would fail to penetrate. Then critical velocity is a continuous random variable. What is left for the experimenter is to characterize the probability measure associated with the random variable, critical velocity. Note that critical velocity is not directly observable since in no way can the experimenter sample directly from a population of critical velocities. Rather, the experimenter can only collect (response, nonresponse) data. If a response is observed at a particular velocity then all that can be said is that that velocity was in excess of the critical velocity for that particular projectile. In this manner data can be collected pertinent to the response function, or the probability distribution of critical velocity. Historically in testing these projectiles, the median of this distribution, V_{50} , is of particular interest primarily because it takes fewer rounds to estimate than other quantiles. We will continue with that convention here.

Our purpose in examining this problem was twofold. The first was to examine the effect of day to day problems in sensitivity testing under a representative 'in practice' scenario. The second was to compare several design and estimation procedures in this 'in practice' setting. Our attention here will be focused on our first purpose.

DESIGN CONSIDERATIONS

A detailed Monte-Carlo study was performed which incorporated some problems encountered in practice. Under each set of test conditions 700 iterations were run giving rise to estimates of V_{50} . The response for this study was taken to be the sample population of the estimate, V_{50} , expressed in terms of the empirical density, its mean, and in particular the $\sqrt{\text{MSE}}$.

The test design appears in Figure 1. Five designs, each in conjunction with three estimation procedures, were used in this study. The Delayed Robbins-Monro (DRM) and the Adaptive Robbins-Monro (ARM) are variations of the well known Stochastic Approximation Method of Robbins and Monro. The Estimated Quantal Response Curve (EQRC), used in conjunction with DRM and ARM in this study, is a recent technique introduced by Wu (1985). The Langlie procedure is one currently used in much of the army's quantal response testing. These five constitute some reasonable designs for use in our testing environment. References are cited at the conclusion of this paper for those interested in the details of these procedures.

The first estimation procedure is a maximum likelihood estimation method with an assumed normal response function and is denoted NMLE. The second (AVR) is an arithmetic average of the velocities giving rise to the k lowest responses and the k highest nonresponses where k is usually taken to be 2 or 3. This second estimate is frequently used by Aberdeen Proving Ground, particularly in the absence of a unique maximum likelihood estimate. The last, Next Stress, is simply the next design point of the sequential design. For DRM, ARM, and EQRC, Next Stress is the intended estimate.

The above designs and estimation techniques were compared under the following test conditions. For some more expensive rounds, experimenters fire 15 rounds in hopes of getting 12 or more. Some are disqualified due to erratic flight of the round. Recently the encouraged policy has been to use as few as 9. Thus, representative sample sizes of 9, 12, and 15 were considered.

Another factor to be accounted for is noise associated with the firing velocity of each round. It is not possible for experimenters to control precisely the velocity at which a round is fired. In fact, for some extensively studied data sets the ratio of the estimated noise standard deviation to the estimated population standard deviation (assuming normal response function) was $.15\sigma$ or more. It was thought that this amount of variation would limit the ability of a sequential design to converge on V_{50} . Three levels of noise were considered: the absence of noise, normal $(0, [.15\sigma]^2)$, and exponential with median, 0, and standard deviation, $.15\sigma$. In each of the above and in the following, σ is the standard deviation of the response function.

Input from the experimenter is used for establishing the initial design point, (starting value) and the range, (gate width) over which the median V_{50} can be found. The latter is used in establishing the magnitude of step sizes in the sequential designs and actually bounds acceptable design points in the case of Langlie's design. Unavoidably, there is often a great disparity between initial estimates and actual values. Consequently, it is reasonable to investigate how well designs and associated estimates rebound from poor initial information. Four starting values were combined with three gate widths in this study.

Finally, it was desired to examine the design and estimator performance under different response functions. Of the five listed only the first four will be considered here. Each have median, 0, and standard deviation, 1, with the obvious exception being the Cauchy whose quartiles were made equivalent to those of the normal.

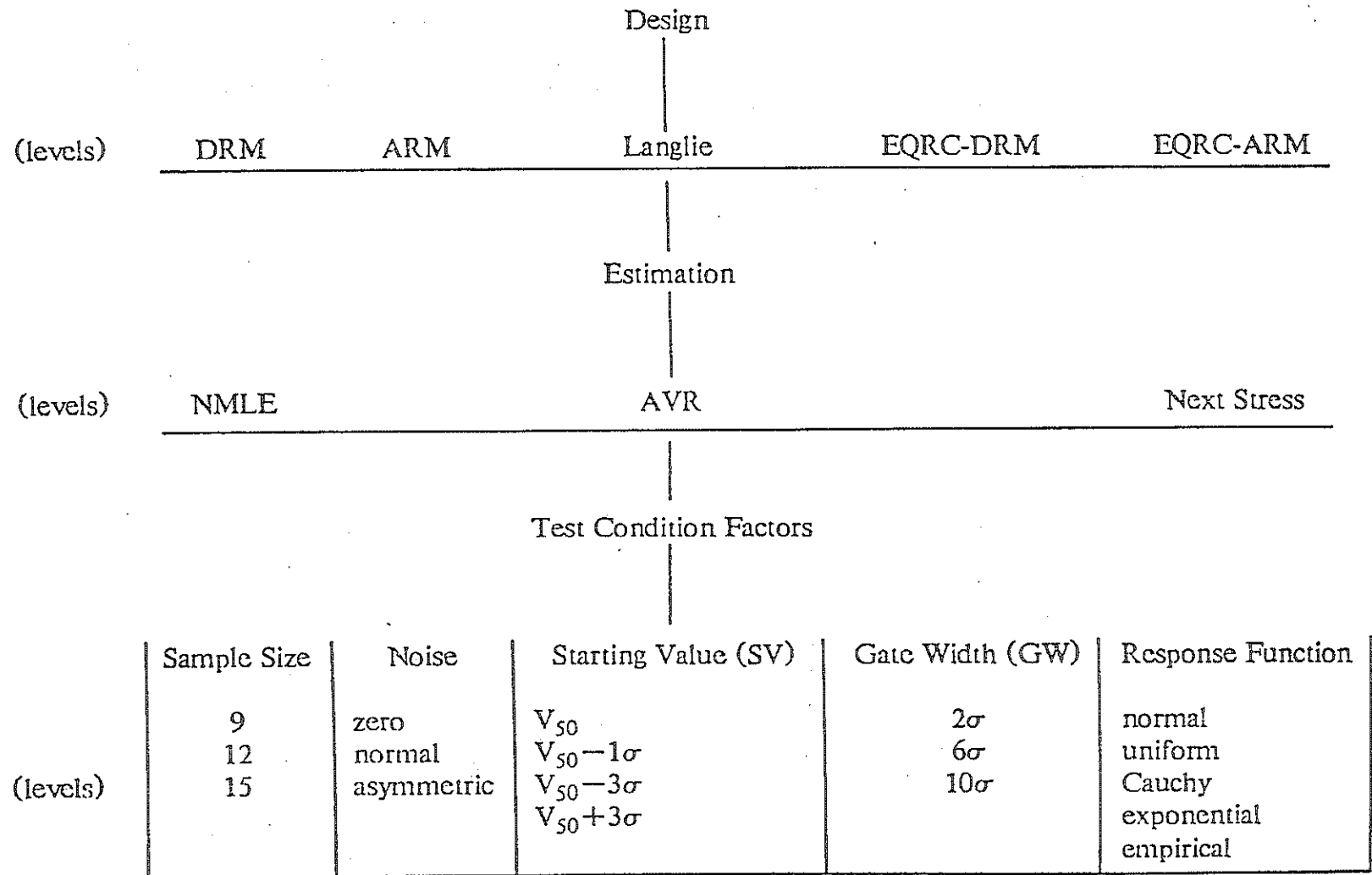


Figure 1. Design and estimation over various test conditions

ANALYSIS

One observation we made was that as the sample size increased, the precision of the estimate improved regardless of the design and estimator used. An example of this is given in Figure 2. We note here that $\sqrt{\text{MSE}}$ is the root mean square error. In addition, a case set is a pairing of a starting value and a gate width. The reader need only know that cases 1-9 are the same in each situation and represent a good mixture of possibilities.

With regard to noise, our study showed AVR and NMLE estimations to be insensitive to normal noise and only mildly sensitive to asymmetric noise. In Figure 3 we see a comparison of \hat{V}_{50}^* 's, the average of 700 simulated \hat{V}_{50} 's for each case set. In the case of asymmetric noise, the average is biased upward slightly toward the longer tail of the response function. However, in Figure 4 we see little difference among the three levels of noise for those same test conditions. We found Next Stress to be sensitive to noise and particularly to asymmetric noise. In Figure 5 the effect of noise on the precision of the Next Stress estimator is evident. In Figure 6 with the actual median indicated by the arrow, note the apparent shift of the estimate population toward higher velocities, the long tail of the asymmetric noise density.

The designs and estimators considered here are influenced by the shape of the underlying response density. In Figure 7 \hat{V}_{50}^* comparisons are made with some zero and normal noise cases. Note that the average of the estimator is approximately the true value of the parameter except in the case of an exponential density and for two cases of the Cauchy density. In Figure 8 these same case sets are compared by $\sqrt{\text{MSE}}$. We see that the uniform density results are somewhat higher than the normal and that the Cauchy and exponential densities each have some extremely low values. This is particularly interesting in the case of the exponential since its estimate population mean was biased upwards. The reason for such behavior rests in the shape of the densities.

Consider for a moment a density with point mass unity representing the critical velocity probability mass. Then if a sequential design were used, the step for the next design point would always be taken in the direction of the point of jump. Thus the design would never make a wrong decision, the decision moving the data collection away from the median. Hence, it would converge in an ideal sense to the median. Of course in order to make a good estimate of the median, it is desirable to sample close to it. Thus, a wrong decision is extremely detrimental over the first few rounds of small sample experimentation as it may prematurely cause sequential designs to decrease step sizes, thus making it more difficult to climb back to the region about the median. For the densities considered here there is a non-zero probability associated with making a wrong decision.

Examine Figure 9. Here all four densities are considered. Suppose for a normal density the sequential design is currently at -2, then we have only a probability of .0228 of making a wrong decision. That is, there is only probability .0228 associated with critical velocities below -2 which would cause a response to be recorded and, consequently, a step down on the stress axis to the next design point. With this in mind, one

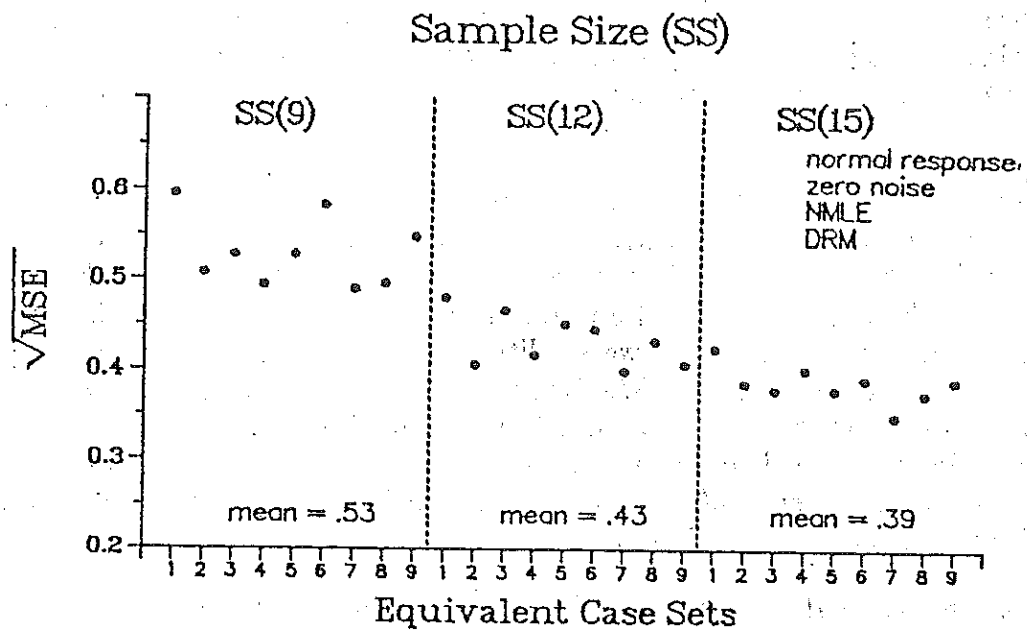


Figure 2. Effect of sample size on precision.

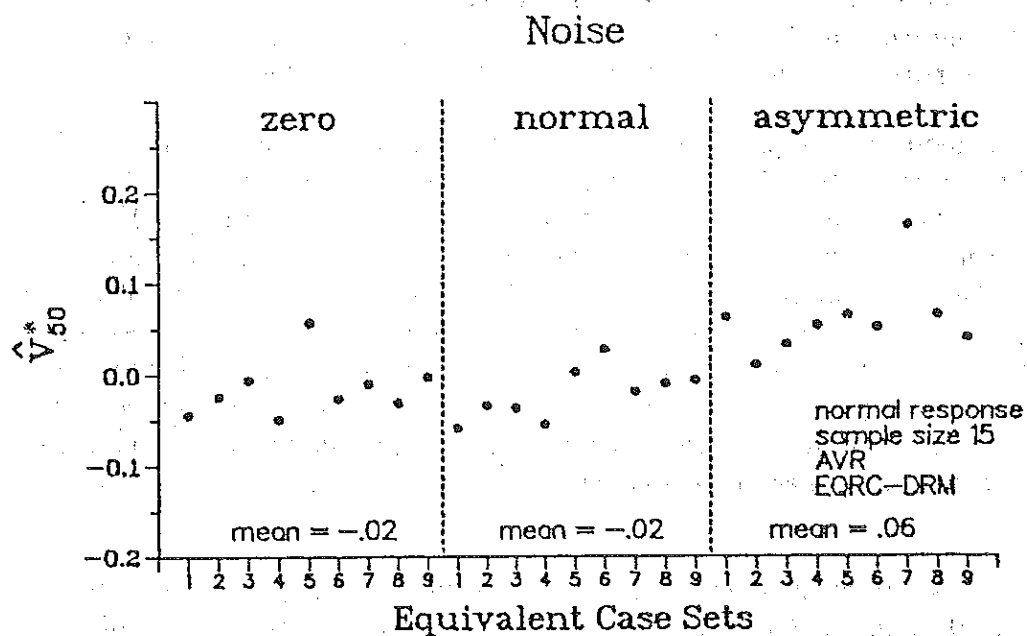


Figure 3. Effect of noise on sample median.

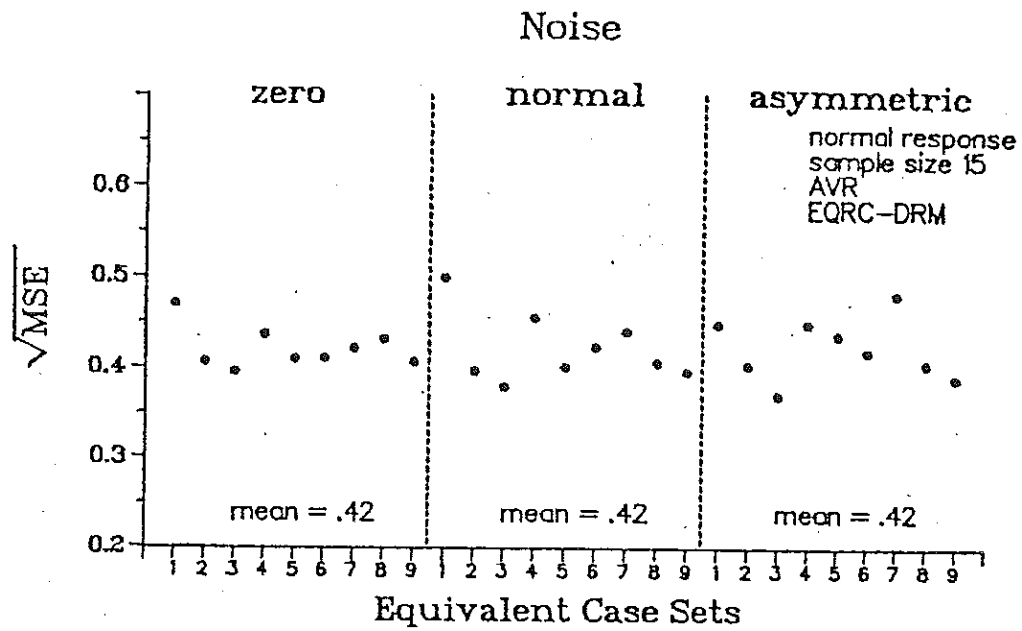


Figure 4. Effect of noise on precision of AVR estimator.

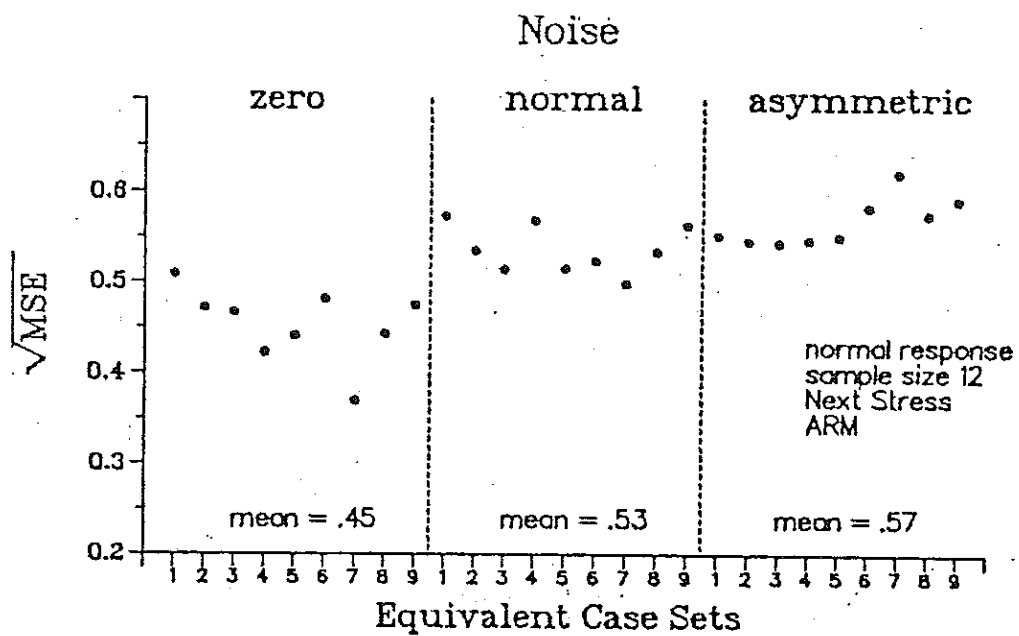


Figure 5. Effect of noise on Next Stress estimator.

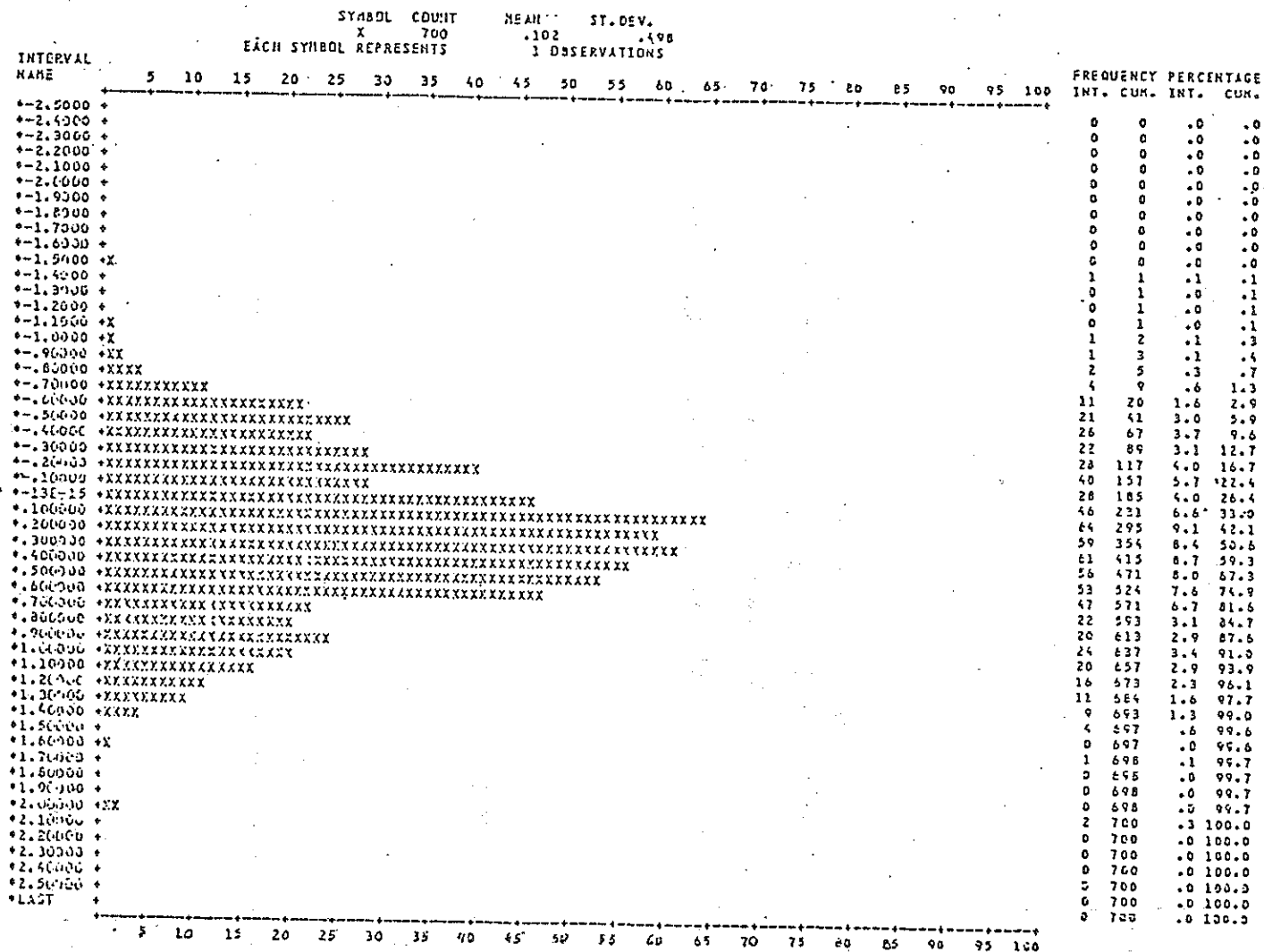


Figure 6. Effect of asymmetric noise on Next Stress

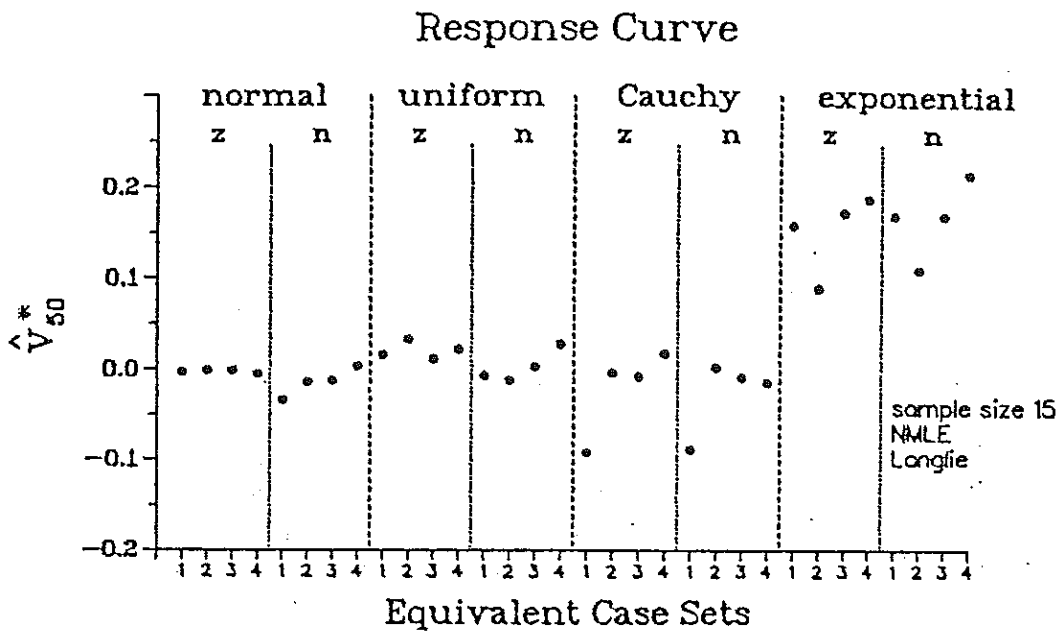


Figure 7. Response curve influence on median estimate.

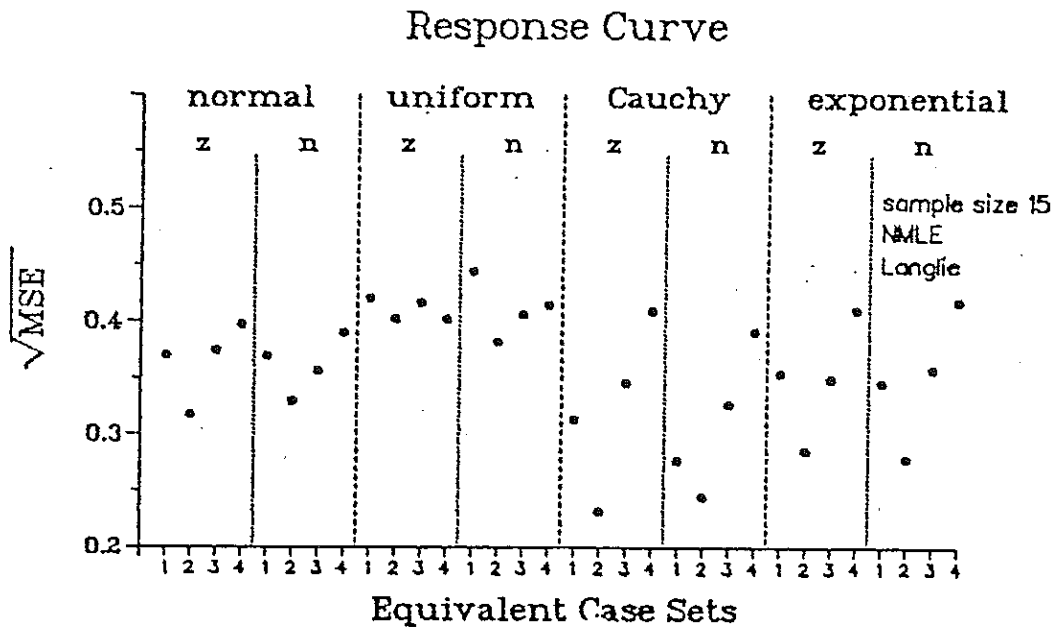
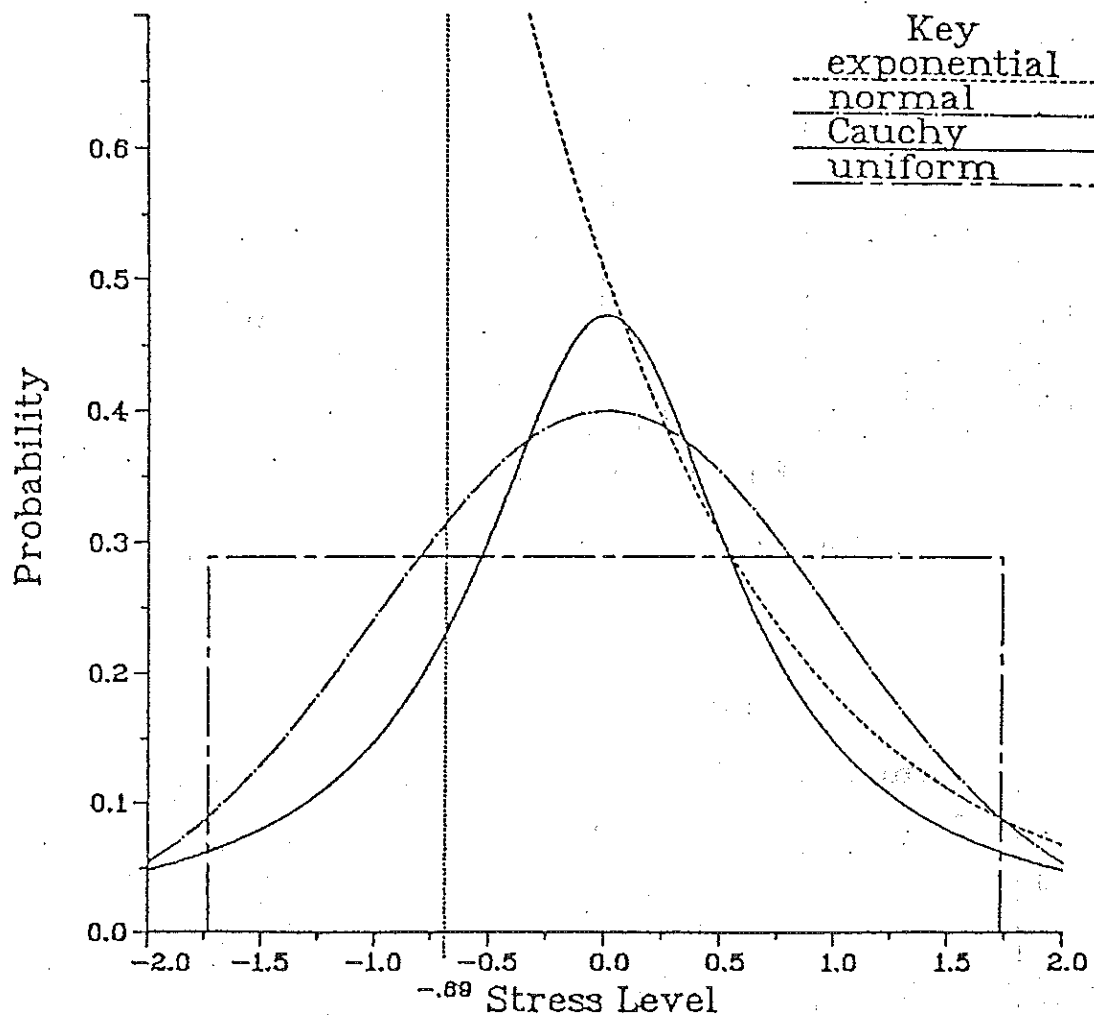


Figure 8. Effect of response curve on NMLE estimator.



can explain the behavior of the designs for each response function.

For the Cauchy density, once the design was sampling close to the median, the concentration of probability in that area was holding the design there. This gave rise to the low $\sqrt{\text{MSE}}$ in Figure 8. On the other hand, for case 2 in Figure 7 where sampling began in the tail, the heavy tail of the Cauchy gave a relatively high probability of going further out in the tail. When the design moved back toward the median, estimation was weighted by the low probability response, resulting in \hat{V}_{50}^* values well below those of the other densities.

In the case of the exponential, most of the probability mass is contained in the interval $(-.69, .69)$ - relatively close to the median. Again, once the design reached this area, the concentration of probability was likely to hold it there, giving rise to Figure 8 results. However, when the design did wander, it could only wander in one direction, thus causing the \hat{V}_{50}^* 's to be higher than for the symmetric distributions. The uniform and normal explanations follow along these same lines.

In support of this explanation we offer as examples Figures 10-13. In each figure the 700 \hat{V}_{50} 's are given in histogram form. Note that -1.1 and 1.3 bound the normal \hat{V}_{50} 's where as -2.5 and 1.8 bound the Cauchy \hat{V}_{50} 's. In addition, the sample estimate population appears slightly more peaked for the Cauchy density than for the normal. Note also the shape of the sample estimate population corresponding to the exponential. It is skewed to the right but at the same time very peaked about the median.

One important idea resulting from these observations rests with the heavy tails of the Cauchy. It is doubtful that with historical small sample data that a normal density could be discerned from a Cauchy with matching quartiles. Yet these simulation results show that problems in estimation can result when heavy tails are present. Therefore, the experimenter needs to be aware of this problem when picking starting values and step sizes.

Thus far only moderate attention has been given to the estimation procedures. In general, we found the NMLE and AVR methods to track very closely over a wide range of starting values and gate widths. Figure 14 shows an example of this in terms of $\sqrt{\text{MSE}}$. However, Next Stress, with its sensitivity to noise environments, does not track well with the other two for normal and asymmetric noise; an example is given in Figure 15. It should be noted that Next Stress is the intended estimator for all designs except the Langlie which uses NMLE. Over the wide range of cases NMLE seems to be the best performer.

The comparison of designs was too involved to address in the time allotted for this talk. We will say only that under NMLE all the designs performed similarly. This is not to say that some are not better than others, but only that in this small sample environment not enough rounds are available to show superiority where it is present.

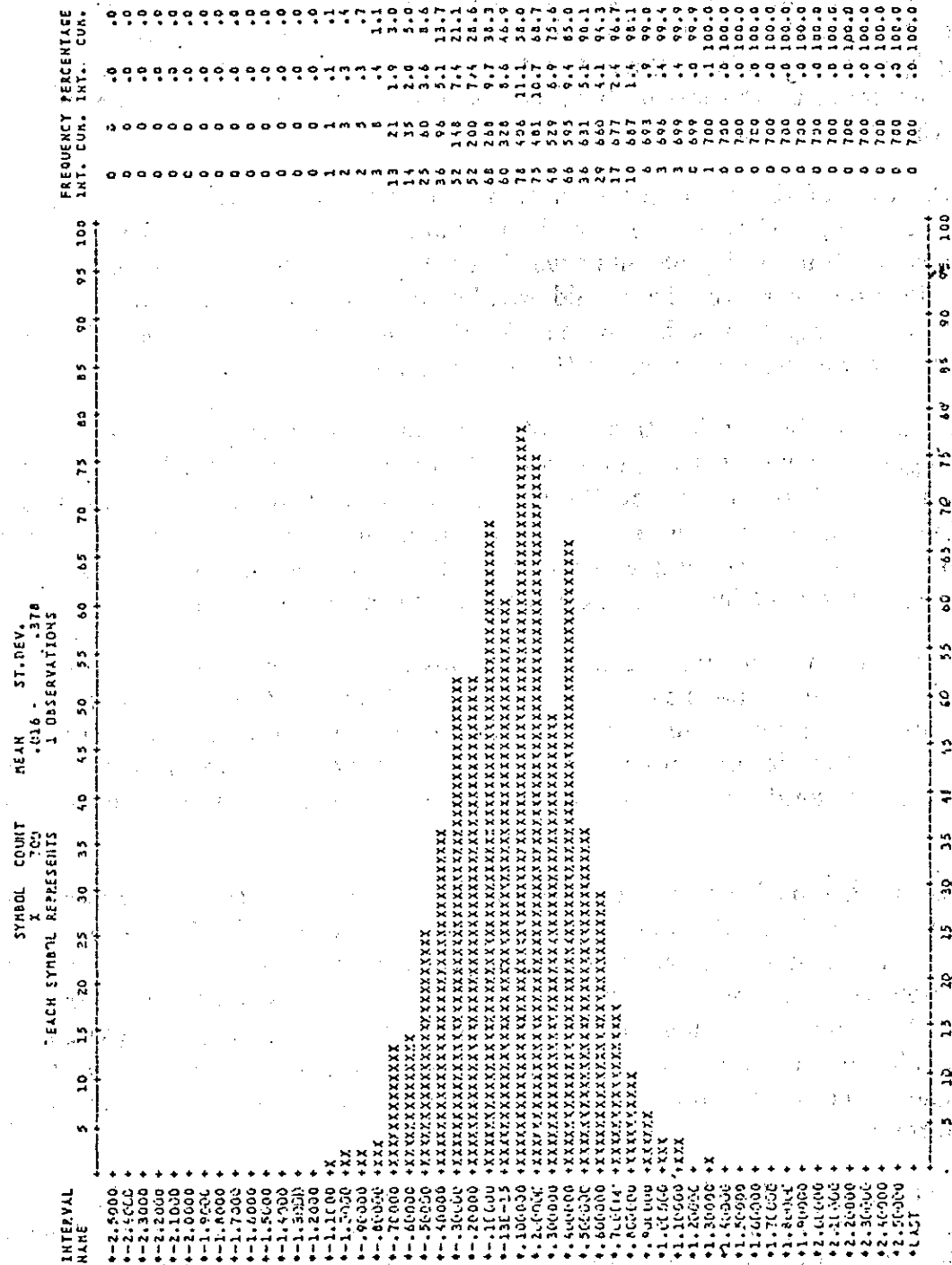


Figure 10. Empirical estimate density for normal response

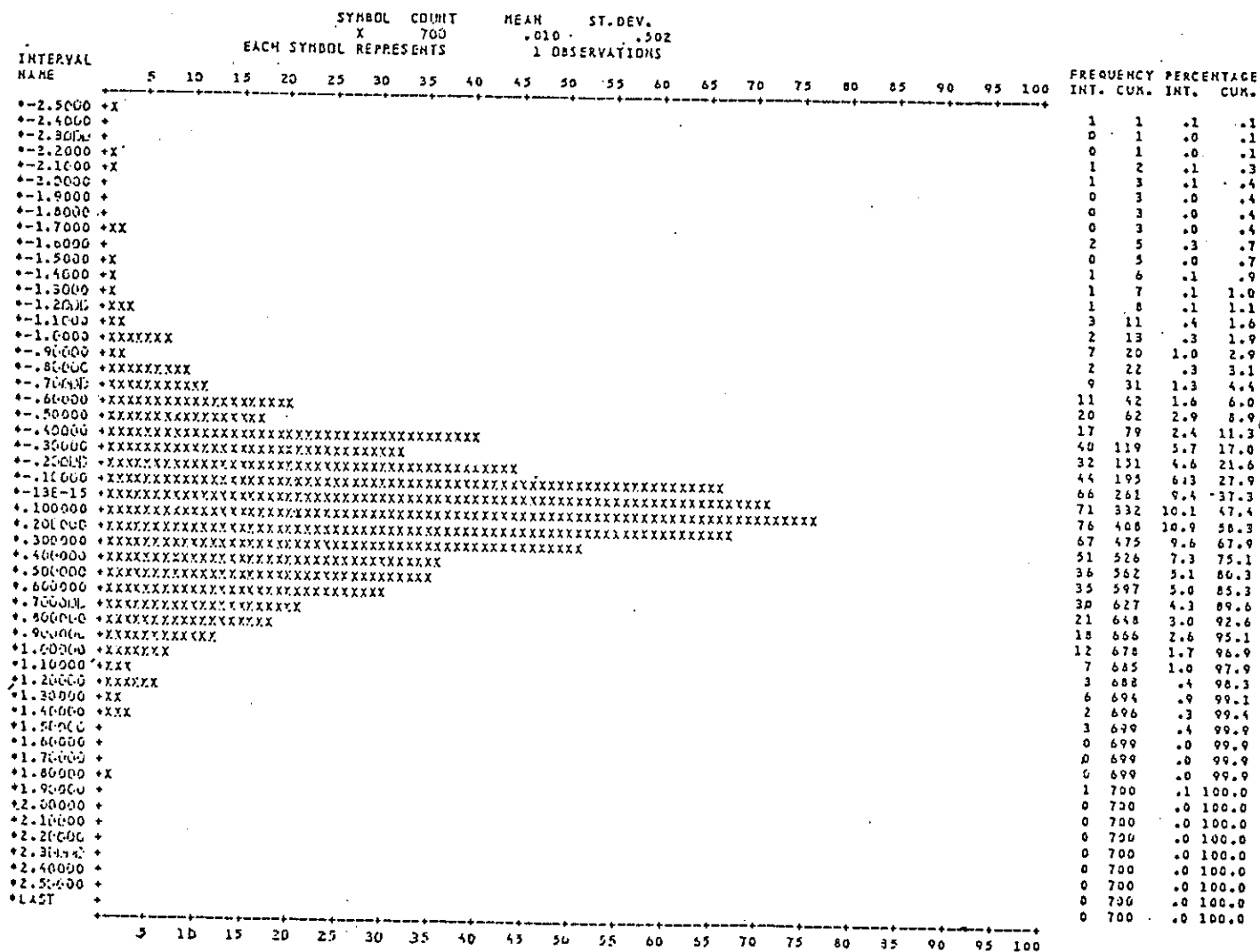


Figure 11. Empirical estimate density for Cauchy response

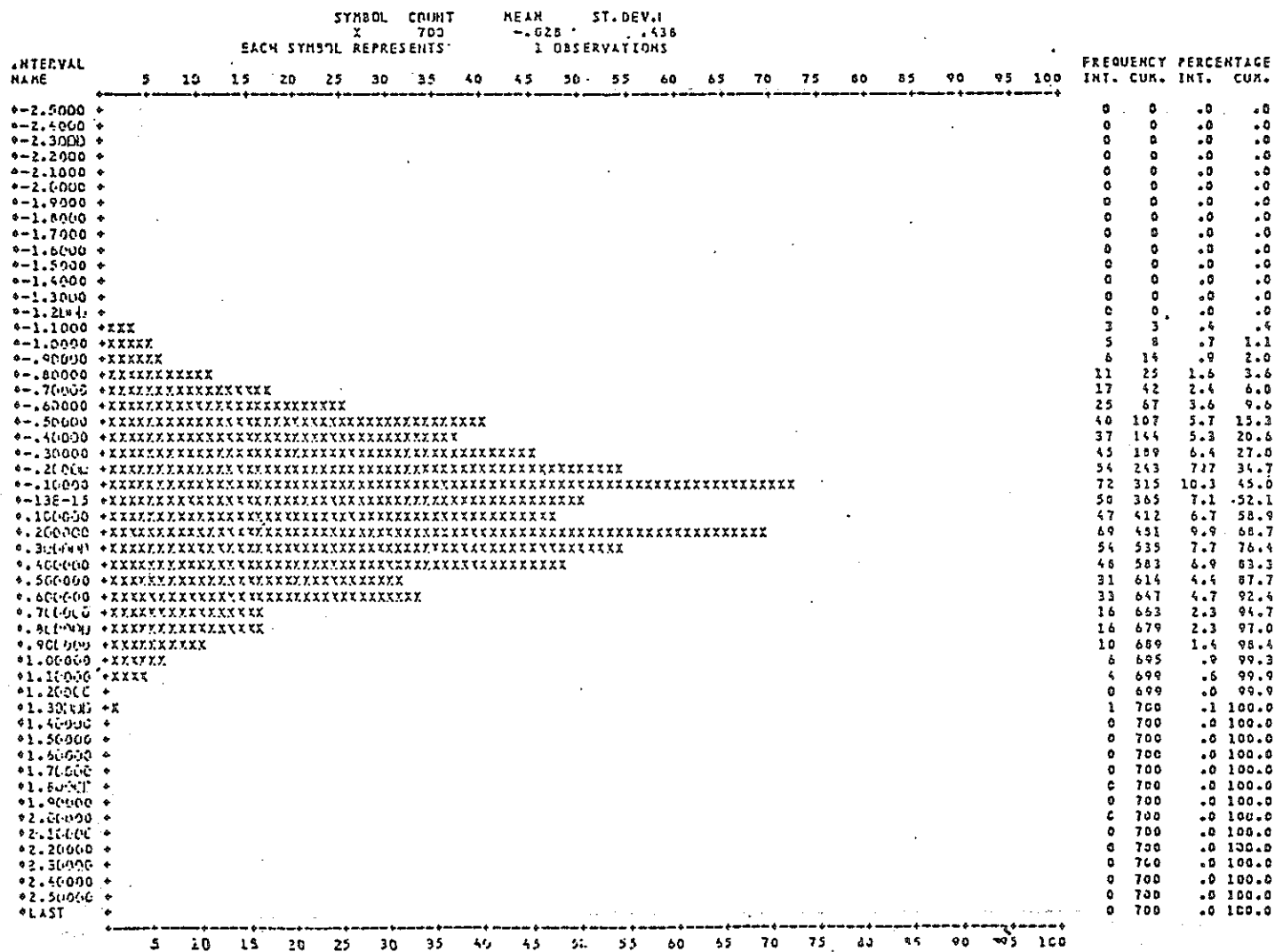
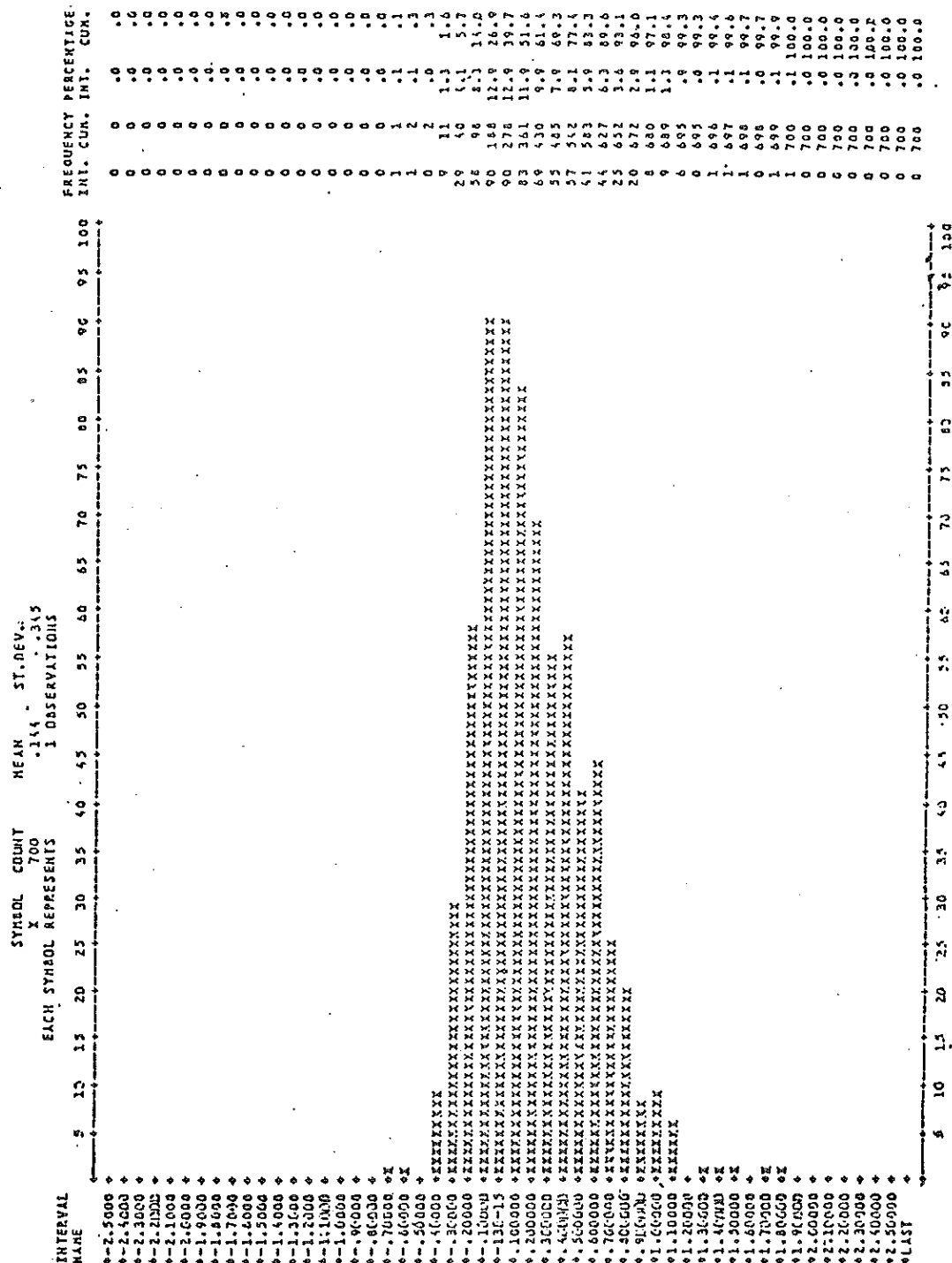


Figure 12. Empirical estimate density for uniform response



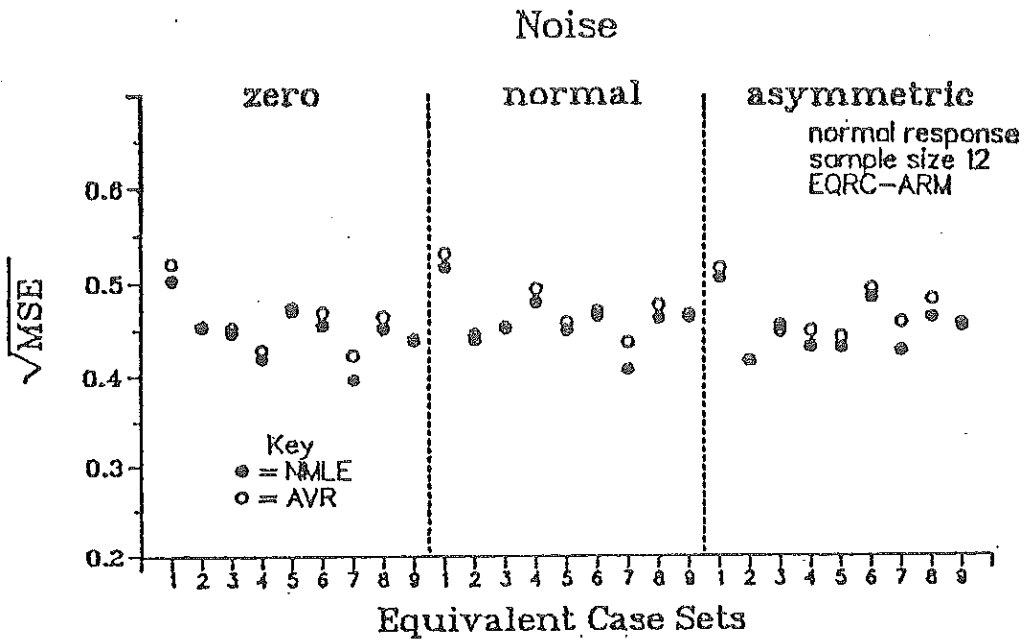


Figure 14. Comparison of NMLE and AVR estimators.

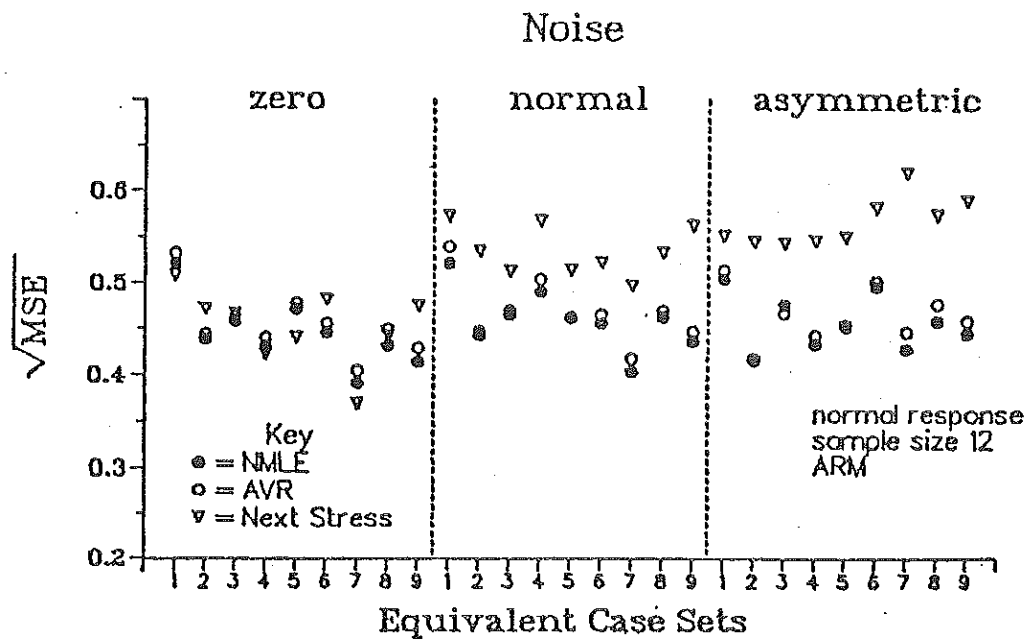


Figure 15. Comparison of NMLE, AVR and Next Stress estimators.

SUMMARY

In summary, several important observations follow. First, the starting value and gate width have a significant effect on $\sqrt{\text{MSE}}$. Second, the response function does influence the design point selection and estimation. In particular, heavy tails could adversely affect the estimate of V_{50} . Third, sample size changes from 9 to 15 result in an increase in precision of about 25%. Fourth, in noise environments, NMLE is the preferred method of estimation regardless of design. In the absence of noise, there is no clear difference among the three estimators. Last, there is no clear advantage in using one design over another in terms of the quality of the estimate. However, certain implementation considerations will help the experimenter choose one to suit his needs.

LIST OF REFERENCES

- Anbar, D. (1978), "A Stochastic Newton-Raphson Method," Journal of Statistical Planning and Inference, Vol. 2, pp 153-163.
- Banerjee, K.S. (1980), "On the Efficiency of Sensitivity Experiments Analyzed by the Maximum Likelihood Estimate Procedure Under the Cumulative Normal Response," Ballistic Research Laboratory Technical Report, ARBRL-TR-02269.
- Brownlee, K.A., J.L. Hodges, Jr., and Murray Rosenblatt (1953), "The Up-and-Down Method with Small Samples," Journal of the American Statistical Association, Vol. 48, pp 262-277.
- Chung, K.L. (1954), "On a Stochastic Approximation Method," Annals of Mathematical Statistics, Vol. 25, pp 463-483.
- Cochran, W.G. and Davis, M. (1964), "Stochastic Approximation to the Median Effective Dose in Bioassay," Stochastic Models in Medicine and Biology, J. Garland ed., pp. 281-300, Madison: University of Wisconsin Press.
- Cox, D. (1970), Analysis of Binary Data, London: Methoen.
- Davis, M. (1971), "Comparison of Sequential Bioassays in Small Samples," Journal of the Royal Statistical Society, Ser. B, Vol 33, pp. 78-87.
- DiDonato, A.R. and M.P. Jarnagin, Jr. (1972), "Use of the Maximum Likelihood Method Under Quantal Responses for Estimating the Parameters of a Normal Distribution and its Application to an Armor Penetration Problem, Naval Weapons Laboratory Technical Report, TR-2846.
- Dixon, W.J. and Mood, H. M. (1948), "A Method for Obtaining and Analyzing Sensitivity Data," Journal of the American Statistical Association, Vol. 43, pp. 109-126.
- Dixon, W.J. (1965), "The Up-and-Down Method for Small Samples," Journal of the American Statistical Association, Vol. 60, pp. 967-978.
- Golub, A. and Grubbs, F.E. (1956), "Analysis of Sensitivity Experiments when the Levels of Stimulus Cannot be Controlled," Annals of Mathematical Statistics, Vol. 57, pp. 257-265.
- Hampton, L.D. (1967), "Monte Carlo Investigations of Small Sample Bruceton Tests," Naval Ordnance Laboratory Technical Report, NOLTR-66-117.
- Hodges, J.L. and Lehmann, E.L. (1955), "Two Approximations to the Robbins-Monro Process," Proceedings 3rd Berkely Symposium, Vol. 1, pp. 95-104.
- Langlie, H.J. (1962), "A Reliability Test Method for 'One-Shot' Items," Aeronutronic Publication No. U-1792.

McKaig, A.E. and Thomas, J. (1983), "Maximum Likelihood Program for Sequential Testing Documentation," Ballistic Research Laboratory Technical Report, ARBRL-TR-02481.

Robbins, H and Monro, S. (1951), "A Stochastic Approximation Method," Annals of Mathematical Statistics, Vol. 22, pp. 400-407.

Rothman, D., Alexander, M.J., Zimmerman, J.M. (1965), "The Design and Analysis of Sensitivity Experiments," NASA CR-62026, Vol. 1.

Rubinstein, R.Y. (1981), Simulation and the Monte Carlo Method, New York, John Wiley & Sons Inc..

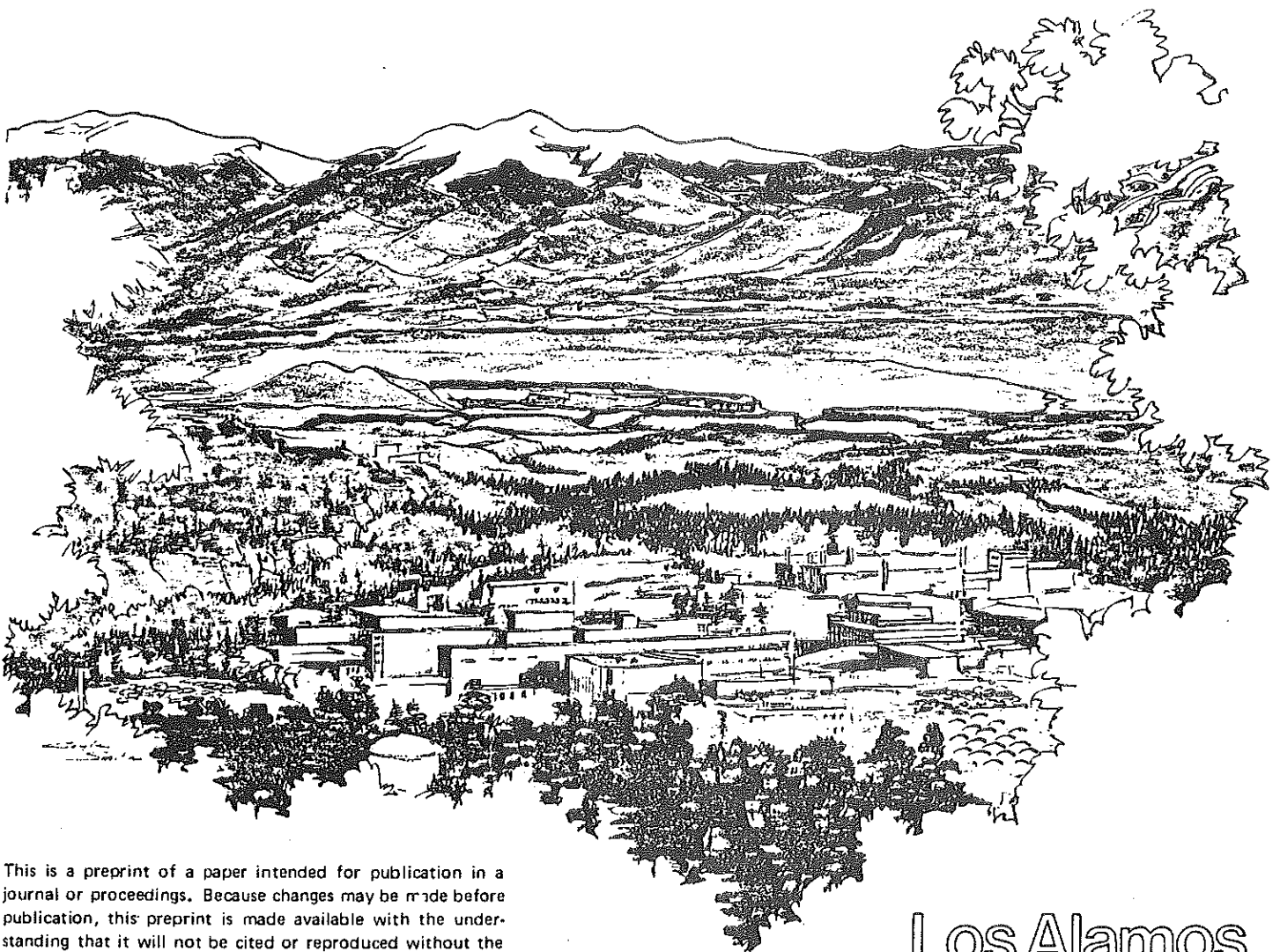
Silvapulle, M.J. (1981), "On the Existence of Maximum Likelihood Estimators for the Binomial Response Model," Journal of the Royal Statistical Society, Vol. 43, pp. 310-313.

Wetherill, G.B. (1963), "Sequential Estimation of Quantal Response Curves," Journal of the Royal Statistical Society, Vol. 25 (1963), pp 1-48.

Wu, C.F. Jeff (1985), "Efficient Sequential Designs with Binary Data," Journal of the American Statistical Association, Vol. 8, pp 974-984.

HUMAN FACTORS AFFECTING SUBJECTIVE JUDGMENTS

MARY A. MEYER
University of California
Los Alamos National Laboratory
Analysis and Assessment Division
P. O. Box 1663
Los Alamos, NM 87545



This is a preprint of a paper intended for publication in a journal or proceedings. Because changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

HUMAN FACTORS AFFECTING SUBJECTIVE JUDGMENTS

Mary A. Meyer
Los Alamos National Laboratory

ABSTRACT

Human factors include the ways in which people acquire, process, and convey information. They affect the quality of people's judgments and thus become a concern when these judgments are being elicited for use as data. This paper focuses on five human factors: question phrasing, conservatism, inconsistency, overoptimism, and social pressures. Techniques for detecting and reducing the occurrence of these human factors are given for two methods of eliciting subjective data, the mail survey and the interactive group method. Techniques for structuring the elicitation methods are proposed as the main means for countering the occurrence of human factors.

THE HUMAN FACTORS

Human factors can affect the quality of the subjective data in many ways. Human factors include the ways in which people acquire, remember, process, and present information that inhibit their reaching mathematically optimal decisions. The human acquisition of data is biased because humans selectively learn that which supports, rather than opposes, their views (Mahoney 1976, Hogarth 1980). For example, people are unconsciously drawn to acquire information which supports, rather than refutes, their preconceptions (Mahoney 1976). Then too, people can acquire faulty information because of the role that feedback plays in the learning process. When people receive no feedback, delayed, or only partial feedback, as often occurs, they may draw incorrect conclusions (Hogarth 1980). For example, scientists who often receive only partial confirmation of their hypotheses are likely to consider this sufficient validation or to believe those data points which support their theory and mentally dismiss the others (Mahoney 1976). The information acquired is stored and may be later accessed by the person during an elicitation session.

How easily such information can be accessed from memory also affects peoples' judgments during an elicitation session. Concrete, catastrophic, or widely publicized information is more easily accessible and thus more greatly influences a person's judgment than less memorable information (Spetzler and Stael von Holstein 1975, Hogarth 1980). For example, it is thought that the League of Women Voters ranked the nuclear industry as posing the greatest occupational hazards to its employees of any industry because of the disproportionate amount of media coverage this industry had received.

The processing of data in the human mind, such as during an elicitation session, is also subject to human factors. Generally, people have difficulty processing more than seven pieces of information at a time (Miller 1956). Typically, they will select a heuristic for solving a problem in a decision situation which then influences the decision they reach. For example, managers may focus on the major aspects of the problem and ignore the uncertainties and complex interactions of factors to reach a decision (Bender et al., 1981). This simplifying heuristic may point to a different decision than one which had included all the complexities of the problem. In applying these heuristics, people are likely to be inconsistent, thus further complicating the gathering of quality subjective data. For example, the manager may have been forecasting the completion date of a large project by adding together the blocks of time that each major phase was likely to require. He may have forgotten to add in a phase being done by a subcontractor, thus failing to consistently follow his own heuristic.

Additional complications may enter as a result of the mode in which participants are requested to give the judgments. For example, respondents may give different judgments on a survey than they would in an interview situation (Payne 1951). They might give varying judgments to different phrasings of the same question (Payne 1951, Sudman and Bradburn 1982, Gorden 1980). Then too, they might give different judgments if they are giving it in "willingness to gamble" or "probability" schemes (Winkler 1967, Hogarth 1980).

Due to the constraints of time, five human factors were selected for discussion. These five factors are widely prevalent and often interrelated as will be described below. The five human factors include the effects of:

- 1) Presentation of the decision task and phrasing of the questions or response options;
- 2) Conservatism;
- 3) Inconsistency;
- 4) Overoptimism and;
- 5) Social pressure.

Evidence of the effect of the presentation of the decision task on the individual's response has been documented by Tversky and Kahnemen (1981). They asked students which alternatives they preferred in gain and loss situations. For example, students chose between: 1) a sure gain of \$250; and 2) a 25% chance of gaining \$1000 or a 75% chance of gaining nothing. In the set of loss alternatives, they chose between; 1) a sure loss of \$750; and 2) a 75% chance to lose \$1000 or a 25% chance to lose nothing. The majority preferred the sure gain in the first pair of options and the risky loss in the second pair. Thus, the relative attractiveness of options varies when the same decision is framed in different ways. Furthermore, individuals are generally unaware of the effect of question framing and, if informed of it, uncertain of how to compensate for its effect.

In addition, there is evidence that the response mode, such as probabilities or equivalent gambles, influence peoples' judgment (Winkler 1967, Hogarth 1980). For example, Winkler (1967) recommended that a "willingness to pay" response mode be used because people gave more conservative, hence more realistic, estimates using this response mode than using probabilities. Similarly, the scales used for the responses, such as 1 to 10 or -5 to +5, can influence peoples' judgments.

The effect of question phrasing has been shown most dramatically by Payne (1951) through his use of the split ballot technique in survey questions. The split ballot technique entails giving half of a survey sample one wording of a question or response option and the other, another. For example, one wording of a question might be, "Do you believe that X event will occur by Y time?" The other wording might be, "Do you believe that X event will occur by Y time, or not?" This second option is more balanced because it mentions both possibilities. For this reason it would be likely to receive a higher percentage of "no" responses. Often the difference measured by the split ballot technique is 4-15% even when the rewording has been very slight.

Conservatism, or anchoring bias, involves the individual's tendency to cling to their first judgment instead of adjusting it to reflect new information. Sometimes this tendency is explained in terms of Bayes' Theorem as the failure to adjust a judgment in light of new information as much as it would be according to Bayes' mathematical formula. Spetzler and Stael von Holstein (1972) and Armstrong (1981) describe how people tend to anchor to their initial response, using it as the basis for later responses. For example, the subject may use the last year's sales as a starting point in predicting this year's sales and fail to consider other points on this distribution independently from this starting point. In addition, Ascher (1978) finds this problem to exist in forecasting where panel members tend to anchor to past or present trends in their projection of future trends. Ascher determined that one of the major sources of inaccuracy in forecasting future possibilities, such as markets for utilities, was the extrapolation from old patterns that no longer represented the emerging or future patterns.

Inconsistency occurs when individuals give contradictory judgments. For example, they might give item A a higher rating than B with respect to goal X, B a higher rating than C, and C a higher rating than A. Inconsistency is a common problem because, as mentioned earlier, individuals are generally unable to apply a consistent strategy, or heuristic, to a series of cases (Hogarth 1980). Inconsistency in an individual's judgment can also stem from his remembering or forgetting information during the process of the elicitation session. For example, the individual may remember some of the less spectacular pieces of information and consider these in making judgments later in the session. Or, the individual may forget that particular ratings are only to be given in extreme cases and begin to give them more freely towards the end of a session than at the beginning.

Overoptimism is sometimes referred to as the overestimation of probabilities, overconfidence bias, or the underestimation of uncertainty. Overoptimism is the giving of more optimistic judgments, such as in the form of probabilities, than the person's data warrants. People tend to be overly optimistic of the probability of some event occurring and often underestimate the uncertainty, or the time and resources needed to make this event a reality. Thus, they give too narrow of error bars on these judgments (Capen 1975). Overoptimism can stem from a variety of causes: 1) thinking at too general a level; 2) wishful thinking; and 3) illusion of control. Armstrong (1975) and Hayes-Roth (1980) have shown that people give higher, less realistic, probabilities when they consider decision tasks in general than when they disaggregate them into their component parts. For example, Armstrong (1975) asked straight Almanac questions of one half of his sample. Of the other half, he asked the same Almanac questions but broken into logical parts. For instance, the question "How many

families were living in the U.S. in 1970?" was asked as "What was the population of the U.S. in 1970?" and "How many people were there in the average family then?". The persons answering the disaggregated questions give significantly more accurate judgments.

Wishful thinking occurs when an estimator's hopes influence his judgment (Hogarth 1980). For example, a project manager in charge of a project may give optimistic probabilities about completing it on schedule because he hopes this will be the case. In general, people exhibit wishful thinking about what they can exhibit in a given amount of time--They overestimate their productivity (Hayes-Roth 1980).

Illusion of control is the tendency to feel greater optimism or greater confidence in some outcome, if one has been involved in its process (Hogarth 1980). People can acquire the impression of having more control over outcomes simply by spending time analyzing a situation as in a elicitation session (Langer 1975). Similarly, people perceive risks as being lower when they feel that they are in control of a process. For example, people perceive less risk when they are driving a car than when they are riding, as a passenger, in a plane (Rowe 1982).

Social pressure induces individuals to slant their responses or to silently acquiesce to what they believe will be acceptable to their group, superordinates, institution, or society in general. Zimbardo, a psychologist, explains that it is due to the basic needs of people to be loved, respected, and recognized that they can be induced or choose to behave in a manner which will bring them affirmation (1983). There is abundant sociological evidence of conformity within groups (Weissenberg 1971). Generally, individuals in groups conform to a greater degree if they have a strong desire to remain a member, if they are satisfied with the group, if the group is cohesive, and if they are not a natural leader in the group. Furthermore, the individuals are generally unaware that they have modified their judgment to be in agreement with the group. One mechanism for this unconscious modification of opinion is explained by the theory of cognitive dissonance. Cognitive dissonance occurs when an individual finds a discrepancy between thoughts he holds or between his beliefs and his actions (Festinger 1957). For example, if an individual holds an opinion which is conflict with that of the other group members and he has a high opinion of the other's intelligence, cognitive dissonance will result. Often, the individual's means of resolving the discrepancy is by unconsciously changing his judgment to be in agreement with that of the group (Baron and Byrne 1981).

Irving Janis's study of fiascos in American foreign policy (1972) illustrates how presidential advisors often silently acquiesce rather than critically examine what they believe to be the group's opinion. This tendency has been called "group think", the "bandwagon tendency", or the "follow-the-leader effect."

The effect of social pressure can also be seen in situations where the individual is not in direct contact with others. Payne (1951) has provided evidence that people give socially acceptable answers to survey questions. On surveys, people claim that their educations, salaries, and job titles are better than they are. More people claim subscriptions to socially acceptable magazines and deny it to the lurid ones than subscription records support. Often there is

a 10% difference between what is claimed for "prestige" reasons and what objectively is.

THE METHODS

Methods for eliciting expert opinion vary along several continuums: 1) the number of participants; 2) the degree of interaction among participants and between them and the session leader; 3) the degree of structure imposed on the elicitation process; 4) the degree of participants' expertise; and 5) the degree of "fuzziness" of the data being elicited.

For example, one method, the mail survey, involves many respondents but little interaction among respondents or between them and an interviewer. Interaction is defined as any two-way communication after which the respondent is allowed to change his judgment. When the respondent fills out a survey, there is generally no interaction between him and his peers or between him and an interviewer.

Another possibility, the Delphi method, can include any number of respondents and allow for more interaction between respondents than the traditional mail survey. The respondents' interactions are controlled by the Delphi monitor who sends each respondent the judgments of the others. The respondents are allowed to adjust their judgments in light of this information. The process of allowing respondents to change their judgments can go through any number of iterations even until consensus is reached. RAND corporation developed the Delphi method to overcome some of the problems inherent in an interactive group method, such as social pressures to conformity. For this reason, in the Delphi technique, the respondents do not interact in a face-to-face situation. Instead, the only contact they are supposed to have with one another is via the mail. And then, the names and other identifying features are removed from the judgments before they are circulated so that the origins of these judgments will not unduly affect the recipients.

Another method, the face-to-face interview, usually involves a fewer number of respondents than the mail survey. The respondents are interactive, singly, with the interviewer during the course of the interview.

Fourthly, there is a interactive group method. In this method, a group of three or more may be convened to give their judgments in the presence of one another. The group sessions are generally monitored and structured by a session leader. For example, the leader may encourage group members to write down their judgments and their reasoning. The leader may require that this information be presented to the group and that a discussion follow. The interactive group method can go through any number of iterations, as in the Delphi method, until consensus, if it is desired, is reached.

For the sake of brevity, this paper will confine its discussion of the detection and reduction of the human factors to two methods, the traditional mail survey and the interactive group method. These two methods were selected because they lie on opposite ends of the continuum with respect to the number of participants and the degree of interaction involved.

The five human factors are manifested in different ways in the various methods so the means by which they can be detected or reduced also vary. For

example, the effect of social pressure is manifested more strongly in the interactive methods such as the face-to-face interview and the interactive group method. Yet, because these methods are interactive, much of the detection of social pressure can be done by a trained observer. This paper's approach to the detection and reduction of human factors in elicitation methods is likely to reflect the orientation of a cognitive or social scientist. The approach is to perform a real time detection or counteraction of the human factors as they occur during a session rather than a later mathematical adjustment of the data.

This paper advocates a structuring of the elicitation methods as a means for reducing the occurrence of human factors. Structuring an elicitation method involves controlling interactions, identifying the parts of the phenomenon on which the respondents are being questioned, defining them and the response options, such as the scale. For example, an unstructured interactive group method would resemble the usual meeting which occurs in the business world. A structured version of the same method would have a program for when each member would present his judgment and rationale to the group, when the floor was open for discussion, and when the next round could begin. In general, the greater the degree of structure imposed on the decision process, the simpler it is to control for the occurrence of human factors. Often a method cannot be maximally structured because each degree of structure imposed slows the process and requires more patience or cooperation on the part of the participants. The client may have deadlines and a fixed budget which limit the amount of structuring which can be done. Thus, the amount of structuring which can be done often involves tradeoffs between the quality of the data and its cost in time and manpower.

The Mail Survey

Detection of Human Factors

In a survey, the occurrence of human factors is not generally detected while the individual is making his judgment but earlier during pilot tests or later when the survey is analysed. Three factors, the effects of question phrasing, social pressure, and inconsistency, can be detected by the use of the split ballot, the sleeper option, and pilot test.

The effects of question wording and sequencing of options can be detected by measuring the differences between the split ballot questions. The split ballot technique is most commonly used for "yes-no" and other multiple choice questions. Use of split ballot techniques in the past (Payne 1951) have shown that people favor generally worded options over those which are highly specific. In addition, they favor options which refer to the status quo over those proposing new alternatives. Split ballot results have also shown that people favor selecting numerical options which are located in the middle of a series whereas they favor nonnumeric options which are located on either end of the series.

Social pressures to give the most acceptable response can also be detected by use of the split ballot technique. One wording on half the surveys can state the options bluntly, the other can contain face saving phrases to encourage people to check the response which is most descriptive of their thoughts or actions. A face-saving option often encourages the respondent to admit that he does not have X knowledge or Y socially-desirable possession at this time by allowing him to state that he plans to acquire them in the future.

Another common area for the effects of social pressures to emerge is in peoples' unwillingness to admit ignorance, to check the "I don't know" option. If identification of knowledgeable respondents is important, a different technique can be used to get a better indication of people's knowledge than simply totalling those who selected the "Don't know" response. A "sleeper" option that sounds plausible but which does not exist in reality can be inserted into the series of bonafide options. For example, on a survey of public opinion of nuclear reactors a "fast water reactor" might be inserted between a "light water", and a "breeder." The number of people who select the sleeper option can be added to those who marked the "Don't know" option and excluded from the pool of supposedly knowledgeable respondents.

Inconsistency in people's responses to surveys is more difficult to detect than the two above mentioned effects. Inconsistency could conceivably be detected by the use of redundant questions but this approach poses problems. If the redundant question is an exact repetition, it can annoy people because they wonder why they are being asked the same question, again. Yet, if the question is asked with a new wording, respondents may give different answers simply because of the difference in phrasing. Inconsistency can occur because the individual has not applied his heuristic consistently, has forgotten instructions or definitions, or has remembered different incidents as he progressed through the survey. An intensive interview type of pilot test can be used to check the survey instrument for these problems. For example, one set of these pilot tests revealed that individuals had forgotten the instructions about half way through the selection of many options. The respondents were supposed to mark their areas of knowledge on a list spanning two pages. Instead by the second page, one fifth of the pilot sample had checked areas in which they would have liked to have had knowledge.

This type of pilot test is the only one, to my knowledge, that can be used to track peoples' thinking, their consistency, through a survey. I adapted several ethnographic interviewing techniques to create this pilot test method. These techniques gather two types of information: 1) how the respondent progresses through the survey, that is which sections he looks at, in what order, and for how long, his general impressions, and when or why he decides to fill out the survey and to turn it in; and 2) how the respondent specifically interprets each direction, question, and response option.

To obtain the first type of information, the interviewee is asked to handle the survey as he would naturally, if no observer were present. The interviewee is asked to "think outloud" and to mention his impressions. Generally, individuals will skim the cover letter and flip through the rest of the survey. As the individual flips through the survey he might state, "I have problems with this page and I would probably let the survey sit on my desk for several days to decide whether to fill it out. While the interviewee pages through the survey, his pauses and gestures, particularly those indicating confusion or anxiety are noted by the monitor. If the respondent has paused or shown some emotion during his review of a particular section, specific questions will be asked such as, "What was your feeling when you read this?"

To obtain the second type of information, the respondent is asked to paraphrase, in his own words, the meaning of each direction, question, and response option. This information allows the monitor to track the respondent's interpretation of each part of the survey.

Structuring the Method to Reduce the Occurrence of Human Factors

As mentioned earlier, structuring any elicitation method can facilitate the counteraction of many human factors. The following section contains some recommendations on how to set up a mail survey to obtain better quality subjective data by controlling for the intrudence of some human factors.

The first stage in developing the mail survey can have an effect on the amount of inconsistency which shows up later in the respondents' judgments. Often seeming inconsistencies in the respondents' answers arise from their viewing the phenomena in a different manner than the way in which it has been presented on the survey. Because the survey does not generally encourage them to explain the view or assumption which allowed them to make the puzzling responses, their responses are dismissed as inconsistent and unreliable. For this reason, it is recommended that the creator of the survey first talk extensively to a sample of those who will be surveyed to learn what relationships, causes and effects, they believe enter into the problem. For example, respondents from a utility might believe that the future of their utilities market is tied to the nation's gross national product (GNP). If the task is to elicit their projections for a utilities market in year 2000, then the questions should define different levels of GNP. For instance, "Assuming that the GNP is X in the year 2000, what would you predict the market for Y to be?"

Careful composition of the questions can reduce the occurrence of three effects: 1) inconsistencies which arise from the respondents' confusion, 2) phrasing, and 3) social pressure. The use of Basic English is recommended if the survey is targeted for the general public as one means for minimizing misunderstandings. Basic English is a vocabulary of approximately 1000 words that are understood by most people who possess a high school education. Payne (1951) provides a list of these words. He also provides a list of words which have been found to possess different meanings for different people. For example, "this year" means the present fiscal year to some, the present calendar year to others, and this coming year to still others. It is recommended that the use of these problem words or phrases be avoided in the interests of clarity. In addition, it is recommended that question lengths not exceed 25 words because respondents' comprehension has been found to fall off around that point (Payne 1951).

As mentioned earlier, the split ballot techniques can be used to detect or counteract the effect of phrasing and ordinality. For example, response options can be placed first or last in half the surveys and in the middle in the other half to counter the effect of ordinality.

If the pilot test of the survey indicated that prestige was an issue on some questions, then face-saving wordings can be used to obtain a better representation of peoples' opinions. Generally, admission of ignorance involves the loss of prestige, so the "Don't know" option should be carefully worded. "No set opinion at this time" is an example of a face-saving wording.

The presence and placement of definitions is another technique which can be employed to reduce the occurrence of human factors, in this case, inconsistency. Definitions include descriptions of the phenomena, the time frame in which the respondent is to consider these, and the scale in which he is to respond. As an individual progresses through a survey, the definitions becomes blurred in his

mind. He relies on his memory of these definitions and often arrives at a working definition which deviates from the original written one. For this reason, definitions should be incorporated into the question or they should immediately proceed it. For example, "What is the probability that the motor generator will reach a maximum power of X for Y amount of time by calendar year September 1, 1984?" The definition of the phenomena has been mentioned as part of the question. The same treatment can be extended to the response scale. For example, the Sherman Kent scale gives these descriptors, "nearly certain", "highly probably", and "We are convinced", to describe a percent ranging from 90 to 99. Both numbers and verbal descriptors, or definitions, are used in attempt to make people mean approximately the same thing when they give the same rating.

Another structuring technique, hierarchically organizing the survey, is helpful in countering the respondents' tendencies to conservatism and overoptimism (Meyer 1982a). Organizing the survey in a hierarchical manner generally entails beginning with specific questions and progressing to more inclusive questions. The respondent is not asked major questions until his memory has been prodded to remember more than just the easily accessible information. Thus, his judgment is not as likely to be anchored to just the first remembered bits of data. Using the hierarchical structure also involves disaggregating questions, as shown in the Almanac example, to counter peoples' tendency toward overoptimism.

The Interactive Group Method

Detection of Human Factors

The effects of phrasing, conservatism, inconsistency, and social pressure can be detected during elicitation sessions by the trained observer who is monitoring this process (Meyer 1982b). Generally, only the presence of these effects, not their magnitude, can be detected by this means. This mode of detection assumes that the group members have been instructed to "think outloud" in interpreting the questions and giving their judgments. (More details on the group members' verbalization of their thoughts will be given in the next section.)

The respondent's verbal feedback on their interpretations of questions allows misunderstandings to be caught during the sessions. Conservatism can also be detected during the session. If an individual continuously holds to his initial judgment, even though there has been a discussion and an opportunity to revise his judgment, he is a likely candidate for conservatism. Inconsistency can be detected when members rate an element differently than they did a comparable one earlier or when their interpretation of a definition appears to change.

The problem of inconsistency arises from more sources in the interactive group method than in the face-to-face interview or the mail survey. This is because the group meetings are held many times whereas the others tend to be one-time deals. Thus, with the usual group method, there is the chance of the members forgetting information, instruction, and definitions over the course of time. One inconsistency which can emerge is the ease with which a response option is applied. For example, the respondents may select the extremes of the scale with varying frequency through time. In general, fatigue during a session seems to contribute to the occurrence of inconsistencies, perhaps because people

are not thinking as carefully. (Fatigue is indicated by briefer responses and by the degree of the participants' horizontal inclination.)

The degree of inconsistency can be detected by use of Bayesian-based scoring and ranking techniques. The group members' judgments can be entered into a scoring and ranking program, such as that of Saaty's Analytical Hierarchical Process, to obtain a rating of their consistency (Saaty 1980).

Social pressures can also be detected by real-time observations. Generally, if consensus is easily obtained, no difference of opinion is voiced, and the group members appear to defer to another member of the group, group think is a strong possibility. Social pressures can come from the members of the group or from the institution sponsoring the decision session. The institution may favor a particular decision outcome and apply pressure on the group members to this end.

Structuring the Method to Reduce the Occurrence of Human Factors

The first stage of the interactive group method, a free association exercise, can be used to counteract the members' tendency toward conservatism. The free association exercise involves having group members mention any and all elements which might have bearing on the phenomena in question. For example, in considering a problem on which technologies should be exported from the United States, some of the major elements a free association might have produced would be the military, economic, political, and technological significance of the export items. The elements mentioned during a free association are usually recorded for the group members to see. Later, the group members will work from these in developing a model of the decision situation. The purpose of the free association exercise is to start with a wide set of possibilities and to narrow these to the pertinent ones. The free association exercise is to counter the human tendency to anchor narrowly on past or present cases which may not hold in the future.

The next stage, the organization of these elements into a model, has bearing on how much inconsistency will be observed when the members are giving their judgments. Highly inconsistent judgments (as determined by ear and by Bayesian techniques) often indicate a need to restructure the model to better represent the members' view. This stage of the method is the most time consuming because the participants are not always conscious of how they mentally model the phenomena. Then too, sometimes they are so conscious of some information that they fail to convey it for incorporation into the model.

The elicitation phase can be structured to include various techniques for countering the effects of social pressure, conservatism, and overoptimism. Perhaps, the most critical of all of the structures placed on the elicitation process is the requirement that participants verbalized their judgments and their reasons for giving such judgments. As mentioned earlier, this verbal feedback allows the method to be monitored for the intrusion of many human factors. For example, if group members appeared to exhibit group think, the method can be structured to promote the opposite bias, conservatism. Groups where conformity is likely to be a problem are cohesive groups, groups where the people have worked together before, or groups where there is a dominating leader (Janis 1972). By requiring group members to write down and then report on their judgments and rationale, they are more likely to get attached to their

judgments and defend them when the discussion begins. I would recommend having each person record and read his judgments before opening the floor to discussion and allowing people to modify their judgments. If there is a strong official or even a natural exoffio leader in the group, that individual should be asked to give his judgments last so as not to influence the other group members. In addition, if there is an official leader of the group, he or she should be encouraged to be nondirective during the meetings. An explanation of the group think phenomena usually suffices to convince them that better discussions and data will result from their refraining from "leading."

If on the other hand, group members appear to be too narrow, or anchoring, in their thinking, a series of extreme scenarios can be introduced for their consideration.

If overoptimism has been detected, the group members can be lead to think in greater detail about the elements of the phenomena. This is done in much the way that the Almanac questions were disaggregated for the survey population.

Another technique, the reviewing of definitions, can help reduce respondents' tendency to be inconsistent because of faulty memory. If at the beginning of every session, definitions are verbally reviewed, members will be more consistent in their definitions through time and between themselves. In addition, each time that their judgment is requested, a statement of the question inclusive of definitions, can be given. For example, "What rating would you give to the importance of element X over Y to reaching goal Z?" Their copy of the scale, in this case a Saaty Pairwise Comparison, should include descriptors or definitions of the ratings.

Another technique for reducing inconsistency is to have the group members monitor their own consistency. For this task, they should have copies in front of them of their judgments, and response scale. A matrix structure of the criteria on which the elements are being judged, the elements, and the judgments work well for this task (Meyer 1982b). Often the group members will view an element in a different light than they did earlier and wish to change the earlier judgment to be in line with their current thinking. If their reasoning does not violate the logic of the model or of the definitions, they should be allowed to make the change. Sometimes, consideration of a new element makes them aware that the model and accompanying definitions did not realistically portray this part of the phenomena. Parts of the original model will need to be changed and some of the process of giving judgments will need to be repeated.

REFERENCES

- Ascher, William (1978), "Forecasting: An Appraisal for Policymakers and Planners," Baltimore: John Hopkins University Press.
- Armstrong, J.S. Denniston, W.B. Jr., and Gordon, M.M., (1975), "Use of the Decomposition Principle in Making Judgments," *Organizational Behavior and Human Performance*, 14, 257-263.
- Armstrong, J.S. (1981), "Long-Range Forecasting: From Crystal Ball to Computer," New York, New York: Wiley-Interscience.

- Baron, Robert A. and Byrne, Donn (1981), "Social Psychology: Understanding Human Interaction," Boston: Allyn and Bacon.
- Bender, Paul S., Northup, William D., and Shapiro, Jeremy F. (1981), "Practical Modeling for Resource Management," Harvard Business Review, March-April, 163-173.
- Capen, E.C. (1975), "The Difficulty of Assessing Uncertainty," Society of Petroleum Engineers AIME 50th annual Fall Technical Conference, Dallas, Texas, September 28-October 1, Paper SPE 5579.
- Festinger, Leon (1957), "A Theory of Cognitive Dissonance," Palo Alto, California: Stanford University Press.
- Gordon, Raymond (1980), "Interviewing: Strategy, Techniques, and Tactics," Homewood, Illinois: Irwin-Dorsey Limited.
- Hayes-Roth, Barbara (1980), "Estimation of Time Requirements During Planning: Interactions Between Motivation and Cognition," Rand report N-1581-ONR, November.
- Hogarth, Robin (1980), "Judgment and Choice: The Psychology of Decision," Chicago, Illinois: Wiley-Interscience.
- Janis, I.L. (1972), "Victims of Group Think: a Psychological Study of Foreign Policy Decisions and Fiascos," Boston: Houghton Mifflin.
- Langer, E.J. (1975), "The Illusion of Control," Journal of Personality and Social Psychology, 32, 311-328.
- Mahoney, Michael (1976), "The Scientist as Subject: The Psychological Imperative," Cambridge, Massachusetts: Ballinger Publishing.
- Meyer, M.A., Peaslee, A.T., Jr., and Booker, J.M. (1982), "A Data-Gathering Method For Use in Modeling Energy Research, Development, and Demonstration Programs," Energy Programs, Policy, and Economics, Florida: Butterworth Publishers.
- Meyer, M.A., Peaslee, A.T., Jr., and Booker, J.M. (1982), "Group Consensus Methods and Results," Los Alamos National Laboratory, report LA-9584-MS.
- Miller, George A. (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," Psychological Review, 63, 81-97.
- Payne, Stanley (1951), "The Art of Asking Questions," Princeton, New Jersey: Princeton University Press.
- Rowe, William D. (1977), "An Anatomy of Risk," New York: Wiley-Interscience.
- Saaty, Thomas L. (1980), "The Analytic Hierarchy Process: Planning, Priority Setting, and Resource Allocation," New York: McGraw-Hill.

Spetzler, C.A. and Stael von Holstein, C-A. (1975), "Probability Encoding in Decision Analysis," Journal of Institutional Management Science, 22, 340-58, November.

Sudman, Seymour and Bradburn, Norman M. (1982), "Asking Questions: A Practical Guide to Questionnaire Design," San Francisco: Jossey Bass.

Tversky, Amos and Kahneman, Daniel (1981), "Framing of Decisions and the Psychology of Choice," Science, 211, 453-58.

Weissenberg, Peter (1971), "Introduction to Organizational Behavior: A Behavioral Science Approach to Understanding Organizations," Scranton, Ohio: Intext Educational Publishers.

Winkler, Robert L. (1967), "The Quantification fo Judgment: Some Methodological Suggestions," Journal of the American Statistical Association, 320, 1105-1120.

Zimbardo, Philip G. (1983), "To Control a Mind," Standford Magazine, Winter, 59-64.

USE OF EXPERT OPINION IN THE RELIABILITY ASSESSMENT OF THE
M1 ABRAMS TANK

BOBBY G. BENNETT

U.S. ARMY MATERIEL SYSTEMS ANALYSIS ACTIVITY
ABERDEEN PROVING GROUND, MARYLAND 21005-5071

1. INTRODUCTION

Modern Army weapon systems tend to be sophisticated, complex, and expensive. The complexity and sophistication are necessary to meet the projected threat and lead to the high cost of both development and procurement. There is also typically an urgency to field the new, more capable equipment as soon as possible. Because of this urgency, the Army has adopted the Single Integrated Development Test Policy wherein government, as well as contractor, testing is utilized to find problems and determine the effectiveness of corrective actions.

The Army acquisition process recognizes that most weapon systems are not mature when subjected to government tests by allowing for reliability growth throughout the development phase. Before proceeding into the production phase, however, there is a requirement to demonstrate that the materiel has achieved the reliability threshold established. Ideally, this demonstration is accomplished by sufficient testing of the final configuration to provide statistically valid estimates. Experience has shown that programs which rely on this technique generally do not achieve the reliability objectives within the allocated resources and time. The second best alternative is to design the tests in a test-fix-test fashion that allows for tracking of reliability by using accepted and proven self-purging reliability growth methodology, such as the AMSAA model. This technique has the advantage of using all test data, thus increasing the applicable sample size over the first alternative, and is successfully used by AMSAA in the reliability evaluation of many Army weapon systems. This technique, in fact, is the preferred technique for assessing reliability at any point in the development cycle. The ability to use this technique, however, is contingent upon several factors, one of which is a requirement to implement the corrective action in a timely manner on the test samples. Unfortunately, it is not always possible to meet the conditions necessary to use the AMSAA Reliability Growth Model, or a similar model, due to the time and money constraints previously discussed; such was the case for the M1 Abrams tank during its Full Scale Engineering Development Phase. In such cases, alternate methods must be used to provide credible estimates of the reliability of the final design at the end of development.

This paper describes the process used to assess the reliability of the M1 Abrams tank, and provides comparisons of these estimates to estimates obtained from later tests of the same configuration. Further, lessons learned during this evaluation are presented along with a brief description of improved and formalized procedures developed by AMSAA in response to these lessons learned.

2. M1 RELIABILITY ASSESSMENT

The M1 Abrams tank had a combat mission reliability requirement of 320 Mean Miles Between Failure (MMBF), to be demonstrated during the Initial Production phase of the acquisition cycle. Recognizing that corrective actions for many of the design faults detected during development test would not be implemented until after test was complete, a threshold of 272 MMBF was imposed on the system to be demonstrated at the completion of the Full Scale Engineering Development (FSED). Early in the FSED testing, it became apparent that the initial design possessed a reliability much less than that necessary to progress into production. With approximately forty percent of the FSED testing complete, the tank was demonstrating an "as-tested" MMBF of 120. "As-tested" MMBF was computed, assuming an exponential distribution, by dividing the total test miles by the total number of failures. At that point in time, although failure analyses had been conducted, very few proposed design changes had resulted in hardware changes on the test samples. In fact, due to the desire to implement corrective action on the test samples as soon as possible, some of the changes to the tank hardware had actually resulted in an increase in total system failure rate and had to be removed. All attempts to fit reliability growth tracking curves were unsuccessful. Since an Army decision review was scheduled shortly, an alternate method had to be considered to assess any growth in design reliability, and to further assess the potential reliability considering proposed, as well as implemented, design changes.

To provide a continuing assessment of the M1 Abrams tank reliability, it was decided to conduct periodic Reliability Assessment Conferences as authorized by AR 702-3. This conference, composed of representatives of the materiel developer, combat developer, development test independent evaluator and operational evaluator, was charged with the responsibility of estimating the reliability of the current configuration and to project the reliability when all identified, but not implemented, corrective actions were taken. In order to accomplish this mission, procedures were developed and agreed to by the conference principals.

2.1 Procedures for Estimating "Demonstrated" Reliability

The term "demonstrated" reliability as used in current Army Regulations has been shortened from what the M1 Assessment Conference termed "reliability adjusted for demonstrated corrective action." Failure rate adjustment for this estimate is made only if there is clear evidence, from representative testing, that a reduction in failure rate has in fact taken place. The following procedure was used by the assessment conference to estimate "demonstrated" reliability:

- ° Establish that design change has been subjected to representative test.
- ° Determine that design change had positive effect on reliability.
- ° Estimate effectiveness of corrective action.
- ° Adjust failure rates and compute adjusted reliability.

2.2 Procedures for Estimating Projected Reliability

The projected reliability estimate allows for adjustment of failure rates for proposed as well as demonstrated design changes. As allowed for in AR 702-3, the combat developer and operational evaluator chose not to participate in this projection, other than offer opinions during discussion. Thus, for the M1 program, projections were made by AMSAA and the M1 Program Manager's Office using the following procedures:

- ° Adjust failure rates for demonstrated corrective actions in accordance with procedures outlined in paragraph 2.1.

- ° Using engineering judgement and experience with similar systems, estimate whether or not proposed change will decrease failure rate.

- ° Using engineering judgement and experiences with similar systems, estimate effectiveness of proposed modifications.

- ° Adjust failure rate and compute projected reliability.

It is evident from the agreed to procedures that significant judgement was inherent in estimation of both the demonstrated and projected reliability. In order to maximize the information available to make this judgement, a requirement was placed on the prime contractor to prepare and provide a document to the assessment conference principals at least two weeks prior to the conference detailing:

- ° Results of failure analyses

- ° Results of all testing (before and after corrective action). If testing was other than on test samples, the contractor was required to detail conditions of test.

- ° Proposed effectiveness factor and rationale.

Upon receipt of the contractor documentation, the AMSAA RAM analyst would provide the information, without the contractor's effectiveness estimates, to engineers with experience in the area of interest and ask the following questions:

- ° Based on the contractor presentation, is there evidence that design change will result in lower failure rate?

- ° What is your estimate of the effectiveness of the corrective action, expressed in terms of reduction in failure rate? Provide rationale.

- ° Could correction of this failure mode result in other failure modes? What, in your opinion, is the most likely failure mode and frequency?

This package would normally be reviewed by three engineers independently. The RAM analyst would assimilate the responses; if in close agreement, the responses would be accepted as appropriate; if not in close agreement, the analyst would discuss the differences with each engineer until the differences

were completely understood or a consensus was reached. The analysts would then discuss the results with his supervisor and they would jointly agree to a position for the conference. This modified delphi approach resulted in a range of effectiveness factors and rationale for discussion at the assessment conference.

The assessment conference was conducted in a democratic process, with open discussion by all principals. A majority vote (3 of 4) was required to consider corrective action demonstrated. If considered demonstrated, the effectiveness factor was then agreed to by voting. Because of the work done at home station, the AMSAA position was normally accepted, particularly if the estimate was close to the estimate provided by the contractor through the Program Manager's Office representative.

2.3 Results of M1 Assessment

The above procedures were used prior to the Army review mid-way through the development test program. At that time, results of the assessment were as follows:

	<u>MMBF</u>
As Tested	120
Demonstrated	145
Projected	256

The demonstrated estimate was not vastly different from the "as-tested" estimate for two reasons; (1) The as-tested estimate included some experience with corrective actions implemented on the test samples and (2) very few of the proposed corrective actions had been tested. Although the tank was demonstrating reliability well below the requirement, a go-ahead decision was granted based on a thorough discussion of the corrective actions identified and the estimates provided by the assessment conference as to the effectiveness of these corrective actions.

These procedures were used during the remainder of the FSED and Low Rate Initial Production test with the following results:

	<u>Mean Miles Between Failure</u>	
	<u>As-tested</u>	<u>Demonstrated</u>
Extended FSED (Phase 1)	234	299
Extended FSED (Phase 2)	308	326
Initial Production (1)	278	351
Initial Production (2)	324	351

- (1) Includes Early Production Process Problems
- (2) Excludes Early Production Process Problems

The configuration of the tank at the beginning of the Extended FSED (Phase 1) was essentially the same as that for which a projected estimate of 256 MMBF was made for the Army review. For all other phases of the test program, the configuration at the beginning of the phase is essentially the same as that

for which "demonstrated" estimates were made during the preceding phase. For example, the estimated value for extended FSED (Phase 2) was 299 MMBF based on Phase 1 testing; the actual as-tested value for Phase 2 was 308 MMBF.

It is of interest to note that the estimated value, in most cases, overestimated the "as-tested" estimate. It was observed that the greatest reason for this was the occurrence of new failure modes, in most part not related to any corrective action. It was also apparent that there had been no provisions in the estimates to account for quality assurance and production process problems inherent in the start-up of a new production process. Historically, this start-up process has resulted in approximately a 10 percent reduction in MMBF.

Overall, the process worked well. Even with the recognized problems, the estimates obtained using expert opinion were within the "statistical noise" of the estimates obtained from further testing of the same configuration.

3. LESSONS LEARNED

Although the estimates obtained by using the procedures discussed were very close to values actually demonstrated later, several problems were noted with the procedures.

- ° There is typically a wide variation in the estimates provided by experts on the effectiveness of proposed corrective action. This paper will not attempt to discuss reasons for this variation, but simply note that it did exist.

- ° Intuitively, it was felt that giving credit for corrective action taken for low failure rate modes resulted in an optimistic estimate of reliability.

- ° The assessment conference procedure allows for control of the conference by the "strong" individual (most persuasive), not necessarily the one with the most knowledge. Estimates arrived at by the conference may thus not have the benefit of the representative input of all experts.

On the positive side, the M1 experience demonstrated that credible estimates can be made using expert opinion, and that low risk decisions can be made in a timely manner without the requirement to test the final configuration for prolonged periods.

- ° The contractors (prime and subs) possess the greatest expertise for the particular design. Contracts must be written to take advantage of this expertise, and in such a manner to allow for significant government interaction, to include the independent evaluators. A conscientious effort is required by the government community, to include use of Government laboratories and independent consultants, to properly assess corrective actions.

4. IMPROVEMENT IN PROCEDURES

The two areas of greatest concern that evolved from the M1 assessment was the uncertainty of the fix effectiveness estimates, particularly for the projected reliability estimates, and the realization that projections were probably optimistic because of giving credit for corrective actions for low

failure rate failure modes without considering the effect of other unseen failure modes. Discussions of these perceptions with personnel from the AMSAA RAM Methodology Office resulted in further investigation of the perceived problems and publishing of several reports to document improved methodology and procedures. Following is a brief synopsis of the published reports with comments on how they may be used to improve future assessments.

4.1 AMSAA Technical Report No. 357, "An Improved Methodology for Reliability Growth Projection", Larry H. Crow, June 82.

In this report, Dr. Crow showed that even when the effectiveness factors are known exactly, the adjusted procedures used in the M1 assessment would still overestimate the system reliability. He further was able to mathematically determine the bias term:

$$B(T) = K h(T),$$

Where K = average effectiveness factor

$h(T)$ = average rate of occurrence at time t
of new failure modes for which corrective
action will be taken

Maximum likelihood methods are used to estimate $h(T)$.

Use of the procedures outlined in this report make it possible to provide an unbiased estimate of system failure rate. The uncertainty in the estimate of the effectiveness factors, however, remained a concern. In order to alleviate this concern, research was conducted on historical fix effectiveness factors and documented in the following report.

4.2 AMSAA Technical Report No. 388, "Reliability Fix Effectiveness for Army Systems", Bruce S. Trapnell, May 1983.

The purpose of this report was to provide a historical data base on fix effectiveness factors for various systems. The advantage to this data base is that it provides a guide to what might be reasonably expected on similar systems, serving as a useful tool to the engineer in assignment of effectiveness factors for projection purposes.

The report details historical effectiveness factors for eleven systems, to include helicopters, tanks, wheeled vehicles and missiles. The average demonstrated effectiveness factor for all systems was approximately 0.70, with relatively small variation.

Work is continuing in this area to determine fix effectiveness by major subsystems, such as engine, electrical system, etc. These data, broke down to subsystem level, will be even more useful for projection for future, more complex systems.

It is recognized that fix effectiveness depends on many factors, and that the past does not necessarily predict the future. The available estimates, however, will provide a starting point and will force the expert to defend large deviations from past experience.

4.3 AMSAA Technical Report No. 399, "Corrective Action Review Team, (CART's)," Bruce Trapnell and Clarke Fox, July 1983.

The purpose of this report is to standardize the procedures for determining effectiveness factors and making projections. It recommends a procedure which uses historical fix effectiveness factor to modify judgmental estimates. It further specifies additional data that must be collected to use the projection model.

5. CONCLUSIONS

Estimates of reliability provided for the M1 Abrams tank using procedures outlined in this paper proved to be quite good, as demonstrated in later testing. To a large degree, the author feels that this is attributed to the expertise of the engineers and analysts involved - and a lot of luck. The procedures could be greatly enhanced by use of available historical fix effectiveness factors and the projection methodology developed by AMSAA. There will continue, however, to be situations where expert opinion will be the prime input to analyses and decisions. It is thus of paramount importance to continue to develop experts and methodology to best use expert opinion.

APPLICATION OF HYPOTHESIS TESTING TO PERFORMANCE APPRAISAL

Richard H. Duncan
Chief Scientist and Technical Director
White Sands Missile Range, NM 88002

and

Paul H. Thrasher
Plans and Quality Assurance Directorate
White Sands Missile Range, New Mexico 88002

ABSTRACT

The Civil Service Reform Act of 1978 mandates performance-based appraisal systems in federal agencies and performance measurements which are accurate and objective to "the maximum extent feasible." In this paper we study two examples in which objectivity can be defined as the establishment of processes which test hypotheses against actual data and the evaluation of attendant a and b risks. In the first example, we use the Poisson distribution to evaluate performance against a standard for courtesy. This model requires that behavior be directly observed 90 percent of the time for acceptably low "rudeness levels" and is thus impractical. In the second example, we propose using the binomial distribution to evaluate the performance of message center clerks who have the task of assigning "Action/Info" and distributing correspondence to elements of a large organization. In this case the amount of inspection required is affordable.

INTRODUCTION

The Civil Service Reform Act of 1978 (CSRA) requires government agencies to establish performance-based appraisal systems under the general supervision of the Office of Personnel Management. In pertinent words of the statute:

Under regulations which the Office of Personnel Management shall prescribe, each performance appraisal system shall provide for establishing performance standards which will, to the maximum extent feasible, permit the accurate evaluation of job performance on the basis of objective criteria (which may include the extent of courtesy demonstrated to the public) related to the job in question for each employee or position under the system.

In compliance with the CSRA, the Department of the Army (DA) established performance-based appraisal systems for Senior Executives (SE), General Merit (GM) employees, and General Schedule (GS) and Wage Grade (WG) employees. Although the three appraisal systems are covered by different regulations and utilize different forms, they share similar structure, vocabulary, and management philosophy to the extent that one may speak of the "Army Appraisal System" (AAS). Under the AAS, supervisors are to provide each employee with a written Individual Performance Plan (IPP) at the beginning of a rating period.

In an IPP, related Tasks/Activities are grouped into Job Elements described by short titles such as Personnel Management, Preparation of Correspondence, Safety, etc. Some Job Elements are mandatory for supervisors; otherwise, a great deal of discretion is allowed in grouping tasks and naming Job Elements. Each Task/Activity is accompanied by a standard which expresses an acceptable level of performance. Additional standards not keyed to specific tasks may be written for the Job Element as a whole. IPPs for supervisors usually involve six to eight Job Elements with several standards per Job Element. Less structure is required to cover a nonsupervisory position.

System doctrine requires that standards be quantified whenever possible, express a range of acceptable performance, and provide the employee an opportunity to excel by surpassing the standards. This doctrine may be breached by the establishment of absolute standards provided such standards are not an abuse of discretion. Absolute standards may be used in situations where a single failure could cause death, injury, breach of security, or great monetary loss. Thus, a standard may require a pilot to make preflight checks 100 percent of the time, but a standard allowing no typing errors would be an abuse of discretion.

At the end of the performance period covered by the IPP, the rating supervisor is required to make an estimate of actual performance (P_i) against each standard (S_i) and make a judgment of Exceeded (E), Met (M), or Not Met (N) for each Job Element. It is common, but sloppy, practice to use the words "exceeded," "met," and "not met" in comparing each P_i to its associated S_i .

These words have been mentioned (selected as names of ratings for entire Job Elements which usually contain more than one standard) and are not logically available for use in any other context.¹ In order to avoid confusion, we use the separate and distinct designators Above Tolerance (A), Within Tolerance (W), and Below Tolerance (B) for this comparison. No algorithm for mapping a (A,W,B) set for a Job Element into E, M, or N is provided in the system design. It is indeed within the purview of a rating supervisor to rate an employee E or M on a Job Element even though a specific P_i to S_i comparison within the element leads to a conclusion of Below Tolerance. (A reviewing official might require that such a supervisor explain his/her decision!) Following determination of the (E,M,N) set of ratings of Job Elements, an OPM approved algorithm is used to arrive at a final adjectival rating of Exceptional (EX), Highly Successful (HS), Fully Successful (FS), Minimally Satisfactory (MS), or Unsatisfactory (U).

So far we have merely provided a brief description of the structure and vocabulary of the AAS. The appraisal systems of other agencies are quite similar. In the remainder of the paper we examine the implications of attempting to be objective within such a system, objectivity being a statutory requirement.

In order to have specific examples, we introduce two mathematical models. In the first we propose to measure courtesy by direct observations. In the second we propose to measure by actual sampling the accuracy of an Action/Info Clerk in an administrative office who is supposed to route incoming mail to the appropriate subdivisions of a large organization. Before continuing, we note that many supervisors write standards in the form "No more

than N substantiated complaints of _____ during the performance period." (The reader may fill in the blank.) For the purposes of this paper we eschew shortcuts which allow conclusions in the absence of data. Instead, we require that actual observations be used to test hypotheses and assess the attendant risks of drawing wrong conclusions. Since one purpose of performance-based appraisal systems is to provide a basis for rewarding employees whose performance is above acceptable standards, the difference between ordinary good performance and exemplary performance should be detectable by the measurement paradigm. Antithetically, less than acceptable performance should also be detectable in order to validate corrective action for nonacceptable performance.

STANDARDS FOR COURTEOUS BEHAVIOR

It is noted in the Introduction that the CSRA specifically mentions "courtesy demonstrated to the public" as an evaluation factor in job performance. In the same legislation, Congress has provided for a suspension of only 14 days or less for four instances of discourtesy within a one year period.² Considerable discussion of courtesy standards has been provided by the U.S. Merit Systems Protection Board (MSPB).³ It is clear from these references that courtesy should not be the subject of an absolute standard. It may seem paradoxical, but a level of rudeness must be allowed if courtesy is to be measured and rewarded. In our own review of IPPs, we note that courtesy standards are commonly imposed on employees in Secretary/Receptionist type positions and rarely on others. As a side comment, this would appear to be unintentional discrimination against incumbents in a particular job category.

We find that courtesy standards are usually written in the "No more than $N \pm \delta$ complaints received" form. We propose a standard written in terms of "No more than $N \pm \delta$ incidents of discourtesy allowed." This would seem to be appropriate since most employees are under direct observation by a supervisor for some fraction of time. (As a thought experiment, we could imagine employing an inspector to observe the employee through a one-way window for whatever fraction of time is needed to ensure objectivity in the sense intended here.) It is assumed that incidents of discourtesy are random, isolated in time, uncorrelated and that the probability of an incident during a time interval is proportional to the duration of the interval. Provided the number of incidents is small, these assumptions are reasonable and permit the use of the well-known Poisson distribution.

Given a "rudeness allowance" of $N \pm \delta$ incidents per year, we can only estimate the actual performance level, P_a , by hypothesis testing. We seek mathematically consistent sets of the following parameters:

F = Fraction of time observed.

R_a = Acceptance range. If the number of observed discourteous acts is within this range, the sample supports the conclusion that the performance is within tolerance with a given risk of being wrong.

α_E = Employee's risk that a within tolerance or better performance will be rated as below tolerance.

α_S = Supervisor's risk that a within tolerance or worse performance will be rated as above tolerance.

β_E = Employee's risk that above tolerance performance will not be detected.

β_S = Supervisor's risk that below tolerance performance will not be detected.

Mathematical details are presented in appendix I-B. A short table of results follows:

<u>Standard</u>	<u>F</u>	<u>R_a</u>	<u>$\alpha_E = \alpha_S$</u>	<u>$\beta_E P_a$</u>	<u>$\beta_S P_a$</u>
2 \pm .5	.90	0-3	.25	1.00 1	.71 3
2 \pm .5	1.00	1-3	.25	.63 1	.65 3
20 \pm 5	.25	1-10	.10	.92 10	.86 30
200 \pm 50	.25	30-73	.10	.18 100	.44 300

In the first line of the table, we set the rudeness level at 2 \pm .5 incidents per year. (The artificiality of setting δ as half of an incident merely facilitates computation in the small N regime.) The proposed fraction of time observed in this line is rather high, 90 percent. Then if the number of observed incidents of discourtesy is in the range 0-3, inclusive, the rater may conclude that performance is within tolerance with a probability greater than $\alpha_E + \alpha_S = .5$ of having drawn the wrong conclusion. It might seem that if the actual number of incidents of discourtesy is 3 against a standard of $N = 2 \pm .5$ the performance was surely out of tolerance. Not necessarily. When $N \pm \delta$ is used to parameterize the Poisson distribution, it applies either to an ensemble of employees, or individual behavior over many performance periods. Then P_a , the actual performance for a given period, becomes a stochastic variable and an observation of three incidents does not show that $N \neq 2$. (Subtleties of interpretation in the small N regime disappear for larger values of N.) The next entry $\beta_E|P_a = 1.00|1$ is the probability (1.00) that a better performance ($N = 1$) would not be detected, and $\beta_S|P_a = .71|3$ is the probability (.71) that a worse performance ($N = 3$) would not be detected. The second line merely exhibits a decrease in risks if inspection is increased to 100 percent. In the final line, we decrease inspection time and lower

risks by degrading the standard to the point of allowing almost four incidents of discourtesy per week. The dilemma is apparent. Objective validation of performance against a high standard requires a lot of inspection time. Maintenance of the objective process with reduced inspection time requires that the standard be degraded to an unacceptable level.

In the case of Callaway versus DA³, the MSPB reversed a removal action against the appellant which was based partially on failure to perform in accordance with an absolute ($N = 0$) courtesy standard. Absolute standards are likely to be judged by the MSPB as an abuse of agency discretion except in "situations where death, injury, breach of security, or great monetary loss could result from a single failure to meet the performance standard measuring performance of a critical element." That issue is quite different from the one addressed here, namely, the objective measurability of performance against a nonabsolute standard.

A standard written in the form "No more than $N \pm \delta$ substantiated complaints of discourtesy during the performance year" has the advantage of being easy to administer. Such a standard places the inspection and reporting responsibility on the public and coworkers rather than the supervisor. However, the measurement is now a joint property of employee behavior and tolerance thresholds of potential complainants. In practice, few or no reports will actually be received. Trivialized and easy to administer standards lead to "Above Tolerance" decisions in the absence of data and contribute significantly to rating inflation. Were it not for the statutory status of courtesy standards, we would recommend that they be used only on a management by exception basis and not ordinarily included in IPPs. The question of whether or not the adoption of this policy would violate the intent of Congress is debatable.

STANDARDS FOR A MESSAGE FORWARDING CLERICAL FUNCTION

The task in this example is that of sorting a large volume of incoming messages, assigning "Action/Info" to each, and distributing the messages to appropriate elements of a large organization. While many actions are purely routine, others require an appreciation of message content and knowledge of the mission and functions of organizational elements. We assume that the workload is sufficiently large to allow use of the binomial distribution to describe sampling without replacement. (The Message Center at White Sands Missile Range processes about 50,000 such actions per year. The function is performed by three to four employees who also have other duties.) We further neglect the fact that "Action" errors are usually more serious than "Info" errors. Performance standards for the employees are assumed to be in the form " $p \pm \delta$ percent of Action/Info determinations are correct." A sample of size n is to be drawn at random for inspection during the performance year. It is assumed that the inspecting supervisor's determination of "correct" or "incorrect" on each sample element is error free. P_a , the actual performance to be estimated, is expressed as a percentage. R_a is the observed range of correct actions within a given sample of size n that allows acceptance of the hypothesis that performance is within tolerance with risks as defined previously. Mathematical details are presented in appendix I-C. As with the previous example, we exhibit a short table of results.

Standard	n	R_a	$\alpha_E = \alpha_S$	$\beta_E P_a$	$\beta_S P_a$
85 \pm 5%	100	76-93	.15	.23 95	.46 75
94 \pm 2%	100	89-98	.15	.60 98	.70 90
94 \pm 2%	500	450-487	.05	.21 98	.54 90
94 \pm 2%	1000	906-970	.05	.02 98	.28 90
94 \pm 2%	1500	1362-1452	.05	.001 98	.16 90
94 \pm 2%	2000	1820-1934	.05	.0001 98	.07 90

In the data selected for presentation, we begin with a pedestrian level of performance, 85 ± 5 percent, a small sample size, $n = 100$, and exhibit rather high risks. As would be expected, the second line shows that escalating the standard and keeping $n = 100$ increases the risks. In the remainder of the table we maintain a high standard and keep increasing n in order to decrease $\beta_E|P_a$ and $\beta_S|P_a$.

We searched for a sample size and risks of about 10 percent or less as exhibited in the last line of the table. An interesting feature of the results is that for fixed $p \pm \delta$ and $\alpha_E = \alpha_S$, $\beta_E|P_a$ decreases much faster than $\beta_S|P_a$ as n increases. Balanced risks of about 10 percent across the board are not inherent in the model. At the sampling level of $n = 2,000$ the risk of being unfair to the employee is negligible. We speculate that competent, self-confident employees would resent increased inspection, although analysis shows that it would be in their best interest. It should also be noted that R_a is wider in every case than the nominal range of $p \pm \delta$ (expressed as decimal fractions) times n . This is to be expected in a stochastic model; observations outside the nominal range do not necessarily indicate an out of tolerance condition. This is not generally understood by supervisors.

Should it turn out that the number of correct Action/Info determinations in the sample of $n = 2,000$ is more than the top of the range, namely 1,934, that fact along with performance against other standards in the employee's IPP should be an evaluation factor in considering the employee for a performance award. Similarly, an observed number of correct determinations below the bottom of the range, 1,820, indicates a need for corrective action. If the scheme is applied to each of three employees, the total sample is $n = 6,000$, about 12 percent of workload. The standard of 94 ± 2 percent is high enough to represent a good operation, yet low enough to allow employees an opportunity to excel. The amount of inspection is affordable and the paradigm is objective.

There is a more sophisticated procedure for making A, W, or B decisions than that given above. Depending on the data, these decisions may be classified as strong or weak. The supervisor may wish to give the employee the benefit of any doubt and escalate a decision from B to W or from W to A, or gain further confidence that a B decision justifies corrective action. The basic theory can be found in the literature of statistics^{4, 5} and an example is provided in appendix II.

A standard relating to filing errors was a second issue in the case of Callaway versus DA.³ No more than two filing errors were allowed during an "annual files inspection." Errors were found during an inspection in preparation for the "1982 Annual General Inspection"; and the agency claimed that the performance standard applied to any inspection. The MSPB thought otherwise and found in favor of plaintiff on this count. One lesson from this case is that inspection related to performance-based appraisal systems should be defined in terms of on-going processes for monitoring performance rather than scheduled general inspections. Moreover, if we may speculate that the filing workload in this case was high enough to allow an analytical model such as the one used in this example, then the standard itself was faulty. It should have been expressed as a percentage of allowed incorrect actions with a range, set high enough to allow a good operation yet low enough to provide the employee an opportunity to excel, and monitored by an objective process. As did the MSPB, we would find in favor of plaintiff, but with different reasoning.

There is another case, that of Walker versus Treasury,⁶ in which the techniques of this paper can be applied in a critique. Walker's task was specifically that considered here, namely distribution of correspondence.

The agency had been operating with an 86 percent accuracy standard. It changed from this rate-type standard to a number of errors-type standard which translates back to a rate standard of $99.5 \pm .2$ percent. Appellant was allowed 0 - 3 errors per month on a workload of about 500 pieces of correspondence. She in fact averaged about 9 errors per month, committed 10 errors during a 1-month probationary period, and was removed from her position. Among other things, she claimed that the new standard was unreasonably high. The agency claimed that other employees were able to achieve the standard, but did not present convincing evidence of this claim to the MSPB. In critiquing this case, we have two findings: (1) The new standard provided no opportunity for any employee to excel. As shown in appendix I, validation of an above tolerance performance would require observation of a negative number of errors, an impossibility. (2) Had the agency wished to document achievability, the table in appendix I shows that the sample size would have had to be larger than the workload, another impossibility. Our analysis is supportive of the MSPB decision to order the reinstatement of Walker to her position.

COMMENTS AND CONCLUSIONS

The Civil Service Reform Act of 1978 mandates performance-based appraisal systems and performance measurement which is objective and accurate to the "maximum extent feasible." It is appropriate, therefore, to systematically investigate the extent to which performance measurement can be made objective and accurate.

In our exploration of this issue, we have chosen examples in which objectivity can be defined in terms of processes which use actual data to test hypotheses and evaluate related α and β risks. This definition of objectivity is a standard tool in all of measurement science. However, in establishing objective processes one also must consider the cost of inspection in time or money. On this basis, the model for validating performance against courtesy standards must be judged impractical, whereas the model for evaluating the work of "Action/Info" clerks in a message center appears to be worthy of adoption.

The analytical approach used in this paper is not applicable in many cases. Some standards are inherently easy to administer. For example, a "Timeliness" standard requires very little inspection time, it being easy to determine whether or not a piece of work is rendered on time. Most performance standards of managers and executives are stated in terms of organizational objectives, do not involve repetitive tasks, and are not amenable to statistical treatment. However, the basic tension between objectivity and inspection time can never be avoided. In this regard, one must also consider the total number of standards to be monitored by a single supervisor. For example, consider a GM-14 who rates three GM-13s and two

nonsupervisory personnel. Job analysis and the structuring of IPPs in accordance with the "school solution" will, in this case, generate about 150 performance standards. Some of these will be easy to administer, some will not. Some will be amenable to hypothesis testing, many will not. In any case, it is clear that effective use of performance-based appraisal systems requires orderly planning of inspection time.

Hypothesis testing should be used in those cases where analysis shows it to be feasible. Any lesser definition of "objectivity" in such cases would be indefensible.

REFERENCES

¹SUPPES, Patrick, (1957) Introduction to LOGIC, Chapter 6, Litton Educational Publishing, Inc.

²Five United States Code, Section 7503(a).

³CALLAWAY, Montine B., versus Department of the Army, (22 October 1984) U.S. Merit Systems Protection Board, Docket No. PH04328310029 Order No. 162.

⁴THRASHER, P. H., (1984), "Modification of Alpha and Beta in Hypothesis Testing," Proceedings of the 30th Conference on the Design of Experiments in Army Research, Development, and Testing, US Army Research Office.

⁵IMAN, R. L., and CONOVER, W. J., (1983), A Modern Approach to Statistics, John Wiley and Sons.

⁶WALKER, Phyllis, versus Department of Treasury (5 Jul 1985) US Merit Systems Protection Board, Docket No. DC04328510014.

Appendix I: APPLICABLE HYPOTHESIS TESTING

A. Background.

Hypothesis testing is a widely used, well documented^s method for comparing a parameter, θ , with a standard, θ_0 . The basic procedure is to assume a null hypothesis, H_0 , and reject H_0 only if there is sufficient experimental evidence that the assumption is unlikely. The significance level, called the Type I risk and denoted by α , is the minimum acceptable likelihood that the experimental data could be obtained if H_0 is true. An alternate hypothesis, H_a , is for use if H_0 is rejected.

The straight-forward hypotheses for performance appraisal would be

$$H_0: \theta_L \leq \theta \leq \theta_U \iff \text{Within Tolerance (W) and}$$

$$H_a: \theta < \theta_L \iff \text{Above Tolerance (A) or}$$

$$\theta > \theta_U \iff \text{Below Tolerance (B)}$$

where θ_0 is replaced by a tolerance range θ_L to θ_U . The Type I risk would be

$$\alpha = P [\text{Rating } \theta < \theta_L \text{ or } \theta > \theta_U \mid \theta_L \leq \theta \leq \theta_U] = P [\text{Rating A or B} \mid W].$$

An opposing risk, called the Type II risk and denoted by β , would be

$$\beta = P [\text{Rating } \theta_L \leq \theta \leq \theta_U \mid \theta < \theta_L \text{ or } \theta > \theta_U] = P [\text{Rating W} \mid A \text{ or B }].$$

The straight-forward way to design the hypothesis test would be to

(1) Select an α .

(2) Select a size for the planned data set.

(3) Use α to determine a range of data, R_a which is defined by $x_A + 1$ to $x_B - 1$, within which a measurement does not indicate a rating of either A or B.

(4) Use $(x_A + 1) < (x_B - 1)$ and the planned data set size to determine β for values of θ such that $\theta < \theta_L$ or $\theta > \theta_U$.

(5) Repeat steps (1) through (4) until the supervisor and employee agree on a triplet of α , planned data set size, and β 's.

Unfortunately, the well-known mathematical relations between α , x_A , x_B , and β 's are based on a standard that is an equality, or at least a semi-infinite range, instead of a finite range. This problem may be handled by performing two hypotheses tests simultaneously. These are:

$$H_0^1: \theta = \theta_L \iff W \text{ or } B$$

$$H_0^2: \theta = \theta_U \iff W \text{ or } A$$

$$H_a^1: \theta < \theta_L \iff A$$

$$H_a^2: \theta > \theta_U \iff B$$

The = signs in the null hypotheses may be replaced with \geq in H_0^1 and \leq in H_0^2 . This change to semi-infinite standards does not change the application of the tests but it does make the interpretation of the tests clearer.

This set of tests will yield a unique member of the A, W, B set for any measurement. The associated Type I and Type II risks are:

$$\alpha_S = \alpha^1 = P [\text{Rating } \theta < \theta_L \mid \theta = \theta_L] = P (\text{Rating A} \mid W \text{ or B})$$

$$\alpha_E = \alpha^2 = P [\text{Rating } \theta > \theta_U \mid \theta = \theta_U] = P [\text{Rating B} \mid W \text{ or A}]$$

$$\beta_E = \beta^1 = P (\text{Rating } \theta = \theta_L \mid \theta < \theta_L) = P [\text{Rating W or B} \mid A] \text{ and}$$

$$\beta_S = \beta^2 = P [\text{Rating } \theta = \theta_U \mid \theta > \theta_U] = P [\text{Rating W or A} \mid B]$$

where the E and S subscripts designate the employee's and supervisor's risks.

The various Type I and Type II risks in the single test and the two simultaneous tests are not as simply interpreted as those for a hypothesis test which has only two possible outcomes. Insight to these relations may be obtained by examining figures 1 through 5. One interesting result, which relates the three Type I risks defined above, is shown by figure 5 to be

$$(\alpha_S + \alpha_E) < \alpha.$$

This inequality can be made to approach an equality only if the actual W domain is made much larger than the domains of B and A.

The two simultaneous hypotheses tests are performed by comparing the measurement, x , with test parameters, x_A and x_B . For a discrete distribution

in which the probability of measuring x events is given by $f(x; \theta)$, the maximum x which implies $\theta < \theta_L$ is denoted by x_A and is the largest x making

$$\sum_{i=x_0}^x f(i; \theta_L) < \alpha_S$$

where x_0 is the lowest value of i making $f(i; \theta) > 0$. Similarly, x_B is the minimum x which implies $\theta > \theta_U$ and is the smallest x making

$$\sum_{i=x}^{x_\infty} f(i; \theta_U) < \alpha_E \quad \text{or}$$

$$\sum_{i=x_0}^{x-1} f(i; \theta_U) > (1 - \alpha_E)$$

where x_∞ is the highest value of i making $f(i; \theta) > 0$. If data yields an x such that $x_A < x < x_B$ or $(x_A + 1) \leq x \leq (x_B - 1)$, the null hypotheses are both accepted and the assumed rating is W . On the other hand, $x > x_A$ implies A and $x < x_B$ implies B .

It should be noted that the calculations of x_A and x_B yield worst case values if the null hypotheses are inequalities. Each equation is the well-known result when the null hypothesis is an equality. The use of θ_L and θ_U as ends of semi-infinite intervals correspond to the worst cases in those intervals.

The Type II risks for the two simultaneous hypotheses tests are calculated from x_A and x_B by

$$\beta_E = \sum_{i=x_A+1}^{x_\infty} f(i; \theta) = 1 - \sum_{i=x_0}^{x_A} f(i; \theta)$$

$$\beta_S = \sum_{i=x_0}^{x_B-1} f(i; \theta).$$

For sufficiently low values of α and large values of $x_\infty - x_0$, β_E and β_S will differ only slightly from the traditional β risk given by

$$\beta = \sum_{i=x_A+1}^{x_B-1} f(i; \theta) = \sum_{i=x_0}^{x_B-1} f(i; \theta) - \sum_{i=x_0}^{x_A} f(i; \theta)$$

because

$$\sum_{i=x_0}^{x_B-1} f(i; \theta) = 1 \quad \text{and}$$

$$\sum_{i=x_0}^{x_A} f(i; \theta) = 0$$

for the values of θ that are of interest in the calculation of β_E and β_S .

The Type I risk, Type II risks, and number of measurements taken are inter-related and competing factors. The balancing of these factors must result from consideration of (1) proposed values of α_E , α_S , and the number of measurements and (2) the mathematically resulting values of β_E and β_S . The employee and supervisor can be aided in their balancing consideration by operating characteristic (OC) curves. The OC-curve is a graph of the Type II risk versus θ with the number of measurements as a parameter. The employee naturally wants an OC-curve with α_E and β_E small while the supervisor wants both α_S and β_S small.

B. Poisson.

The Poisson distribution function,

$$p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \dots,$$

describes the distribution of the random variable x in time t provided that t can be divided into intervals Δt such that:

$$i) P [x > 1 \text{ in } \Delta t] = 0,$$

$$ii) P [x = 1 \text{ in } \Delta t] = (k) (\Delta t) \text{ where } \lambda = kt, \text{ and}$$

iii) x_i is independent of x_j where i and j refer to any two different intervals.

The use of $f(x;\theta) = p(x;\lambda)$, $\theta = \lambda$, $x_0 = 0$, and $x_\infty = \infty$ in the equations of Section I-A yields formulas for the design of simultaneous, Poisson hypotheses tests to select an A, W, or B performance rating.

The parameter λ is a meaningful property to test. It is the mean value of x in time t . (Interestingly but usually less directly applicable, λ is also the variance of x in time t .) The additive property of λ ,

$$\lambda_{t_1+t_2} = \lambda_{t_1} + \lambda_{t_2}$$

for any nonoverlapping times t_1 and t_2 , makes the actual substitution for the parameter θ equal to the product of F and λ instead of λ . Here F is the fraction of the time t , for which λ is the mean, that observations are made in the measurement of x .

C. Binomial.

The binomial distribution function,

$$b(x;n,q) = \frac{n!}{(n-x)! x!} q^x (1-q)^{(n-x)} \text{ for } x = 0, 1, \dots, n,$$

describes the distribution of the random variable x provided the following conditions are met:

i) x is the number of "bad" events in a random sample of size n selected from an infinite, dichotomous population.

ii) $P [x = 1] = q$ when $n = 1$.

The use of $f(x; \theta) = f(x; \theta) = b(x; n, q)$, $\theta = q$, $x = 0$, and $x_{\infty} = n$ in the equations of section I-A yields formulas for simultaneous, binomial hypotheses tests to select an A, W, or B performance rating.

Either the parameter q or its mirror image parameter $p = 1-q$ is a meaningful parameter to test. They are respectively the fraction defective and fraction correct of the population. To use the language of "goodness" instead of "badness", simply substitute $1-p$ for θ and $y = n-x$ for x and use $y_A = n-x_A$ and $y_B = n-x_B$ in the acceptance range of $y_A > y > y_B$. When either the p or q description is desired, it may be advantageous to do the calculations in the opposite interpretation because of available tables and/or computer programs.

The design of a binomial hypotheses tests involves the balancing of α , β_E , β_S , and n for a justifiable tolerance interval. Figures 6 and 7 present OC-curves for a reasonably high tolerance interval and low Type I risks. These may be used to balance the risk and the amount of data taken.

Another example, with an inordinately high tolerance interval, is summarized in the table below. The standard used is $99.5 \pm .2$ percent "goodness" or $q_U = .003$ and $q_L = .007$. The Type I errors used are $\alpha_E = \alpha_S = .05$. The last two columns present two points of the OC-curves.

n	x_A	x_B	y_A	y_B	R_a	$\beta_E p_a$	$\beta_S p_a$
500	-1	8	501	492	493-500	1.00 .9985	.93 .9915
2000	1	21	1999	1979	1980-1998	.80 .9985	.81 .9915
6000	10	54	5990	5946	5947-5989	.29 .9985	.64 .9915
18000	41	146	17959	17854	17855-17958	.004 .9985	.27 .9915
36000	90	279	35910	35721	35722-35909	.000003 .9985	.06 .9915

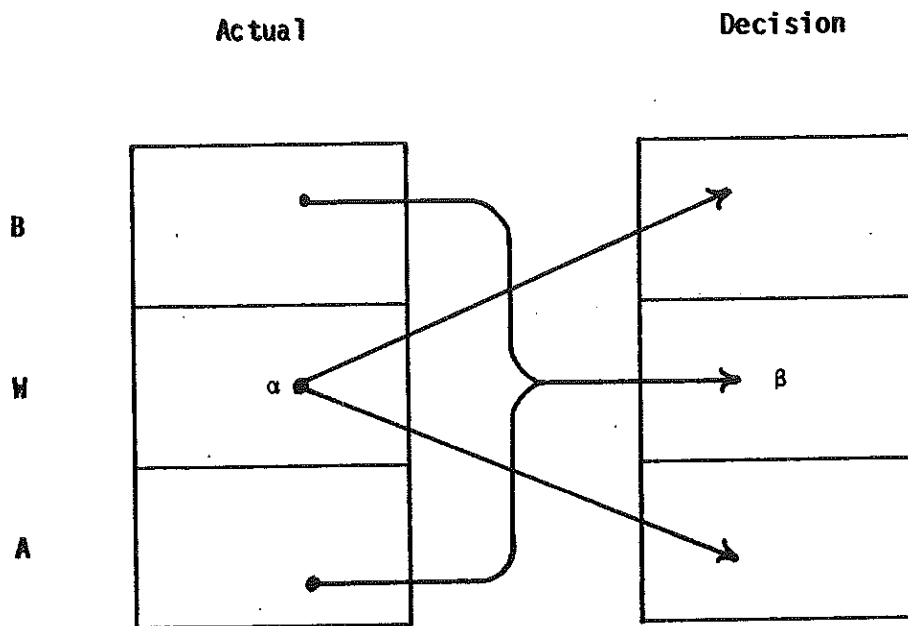


Figure 1: Transitions from actual conditions to rating decisions when one hypothesis test is used. Horizontal transitions would have no risks. Risks of changing B, W, or A are labeled with the appropriate Type I or Type II risks.

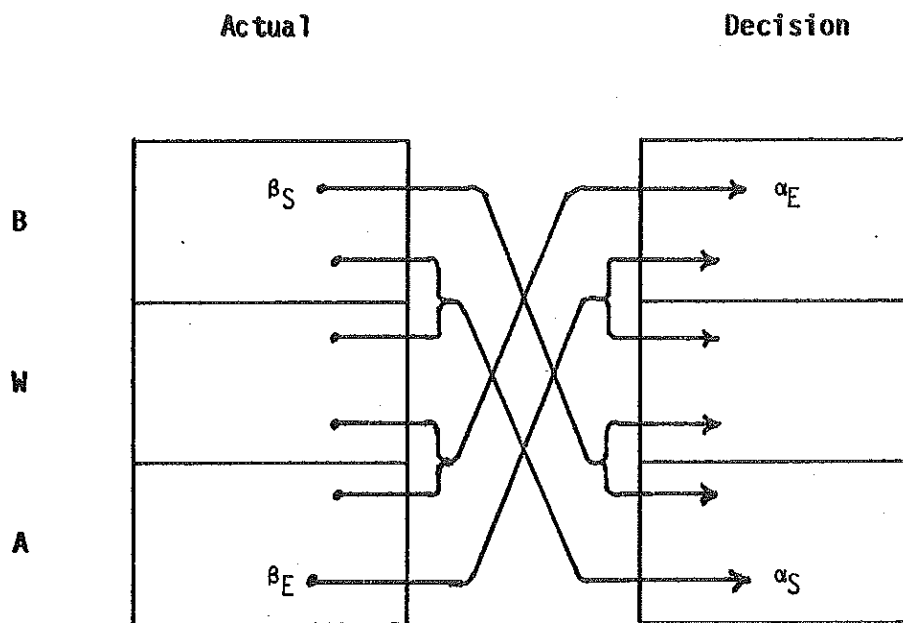


Figure 2: Transitions from actual conditions to rating decisions for two hypothesis tests. Horizontal transitions would have no risks. Risks of changing B, W, or A are labeled with the appropriate Type I or Type II risks.

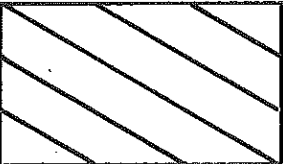

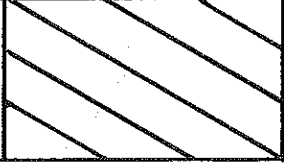
		Actual		
		B	W	A
Decision	B		α	
	W	β		β
	A		α	

Figure 3: The nine possible combinations of actual conditions and rating decisions as viewed with one hypothesis test. The three blocks with downward to the right shading represent correct decisions and have no associated risks. The four blocks labeled with α and β represent risks that are covered by the indicated Type I or Type II risks. The two blocks that are unshaded and unlabeled have risks that are not addressed by the test.

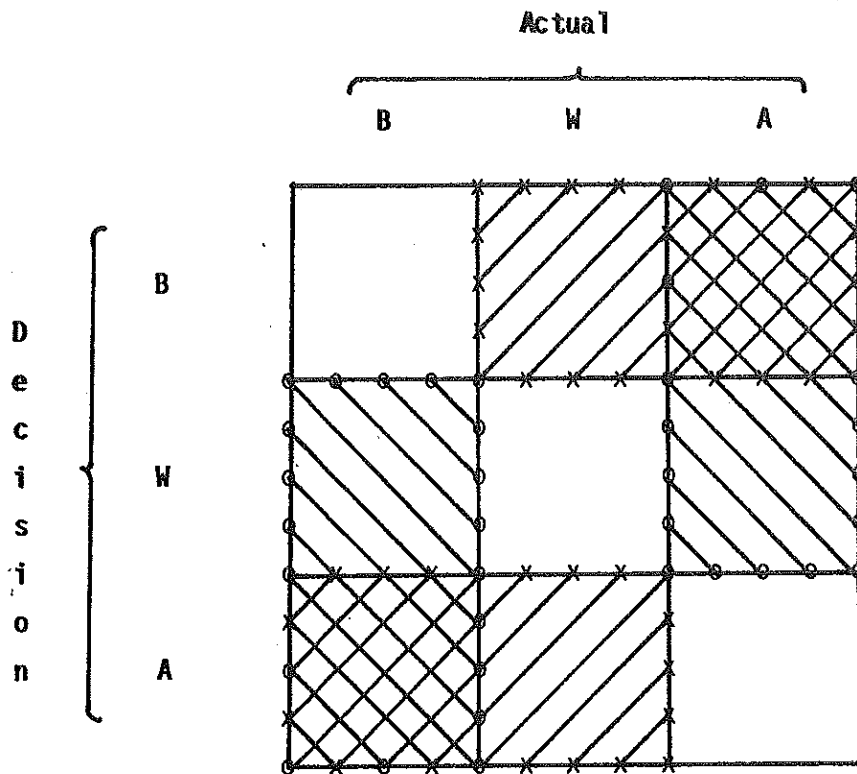


Figure 4: The nine possible combinations of actual conditions and rating decisions as viewed with two hypothesis tests. The three unshaded blocks represent correct decisions and have no associated risks. The three blocks toward the upper-right have associated employee risks because the decision is lower than actual conditions. Conversely, the lower-left blocks have supervisor risks. Shading that is upward to the right indicates that the block is covered by a Type I risk. Conversely, downward to the left shading indicates a Type II risk. Note that two blocks are double covered.

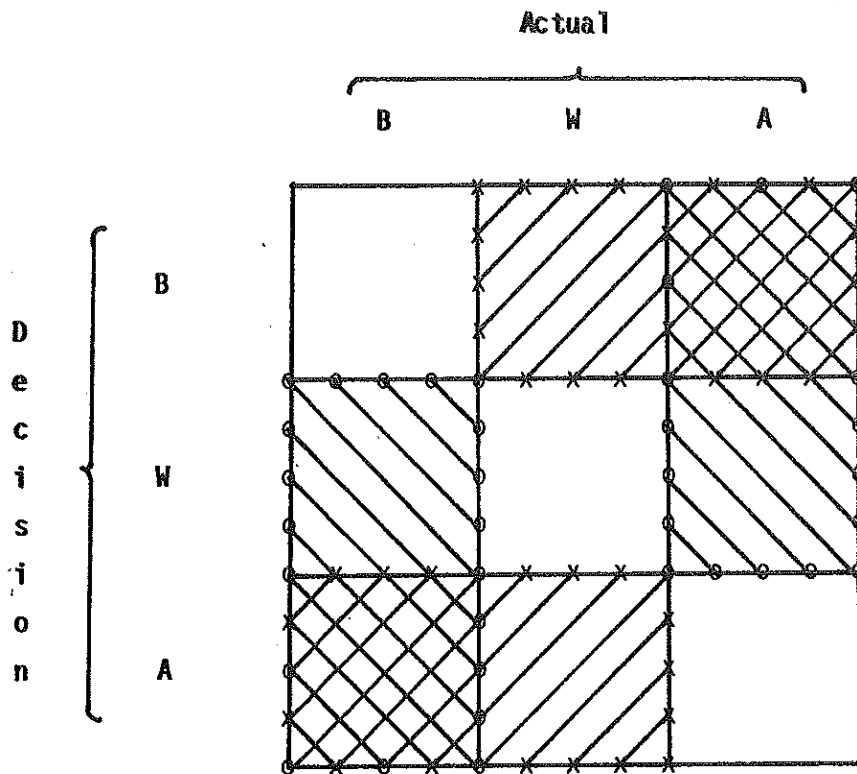
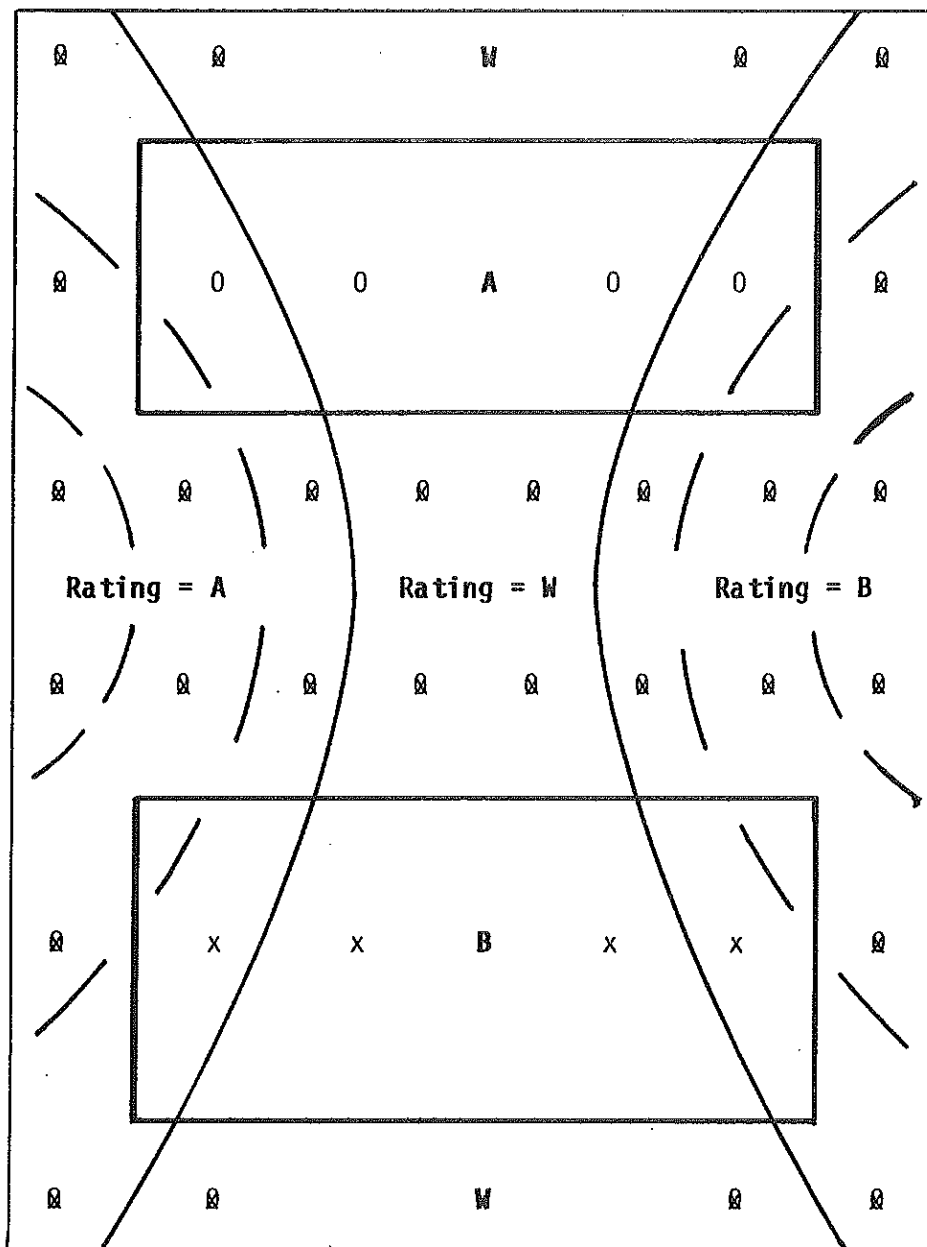


Figure 4: The nine possible combinations of actual conditions and rating decisions as viewed with two hypothesis tests. The three unshaded blocks represent correct decisions and have no associated risks. The three blocks toward the upper-right have associated employee risks because the decision is lower than actual conditions. Conversely, the lower-left blocks have supervisor risks. Shading that is upward to the right indicates that the block is covered by a Type I risk. Conversely, downward to the left shading indicates a Type II risk. Note that two blocks are double covered.



$$\alpha_S + \alpha_E = \frac{)))}{0 + X} + \frac{(((}{0 + 0} < \frac{)))}{0} + \frac{(((}{0} = \frac{))) + (((}{0} = \alpha$$

Figure 5: Venn diagram showing relation between Type I risks.

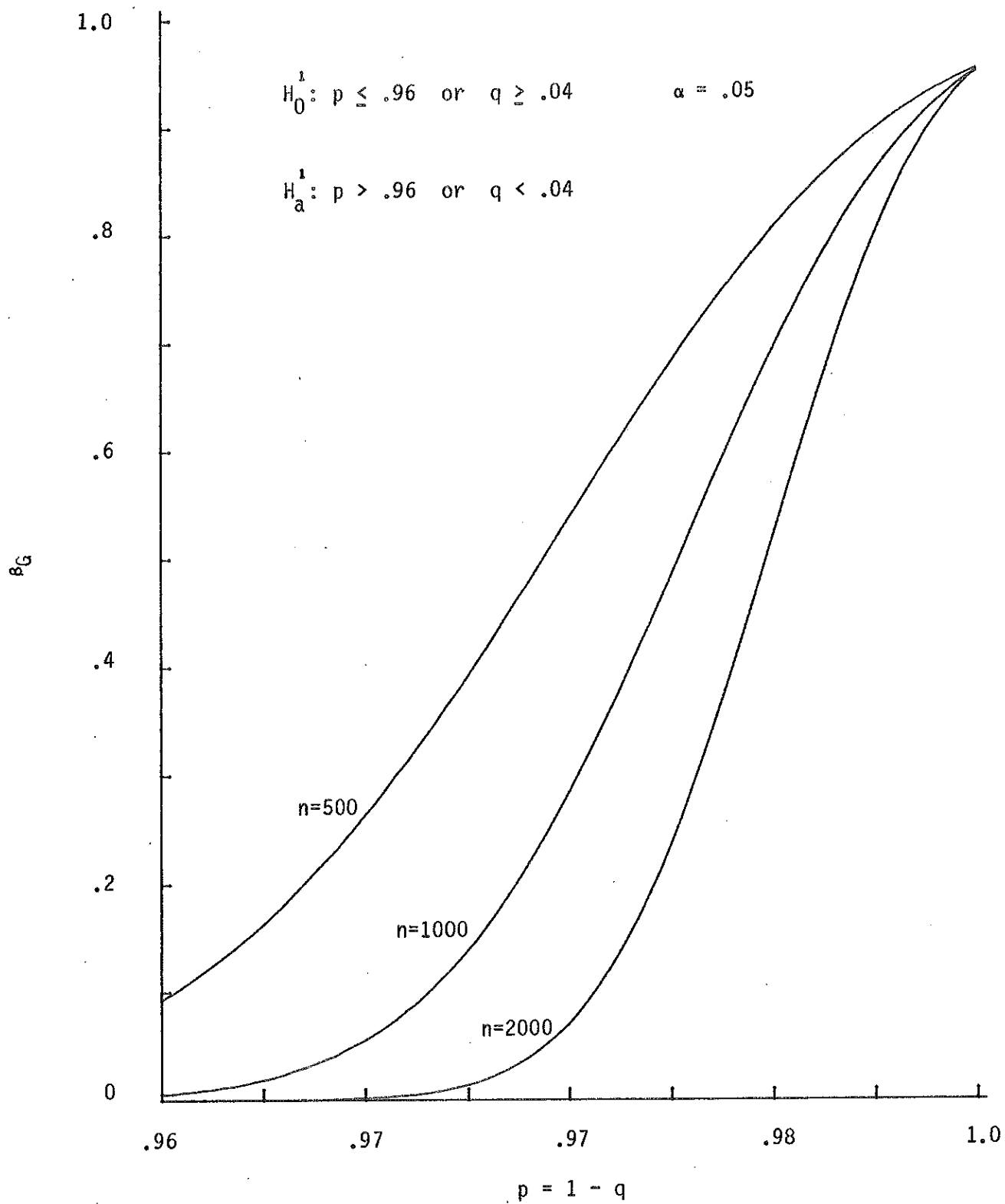


Figure 6: OC-Curve for Upper Test

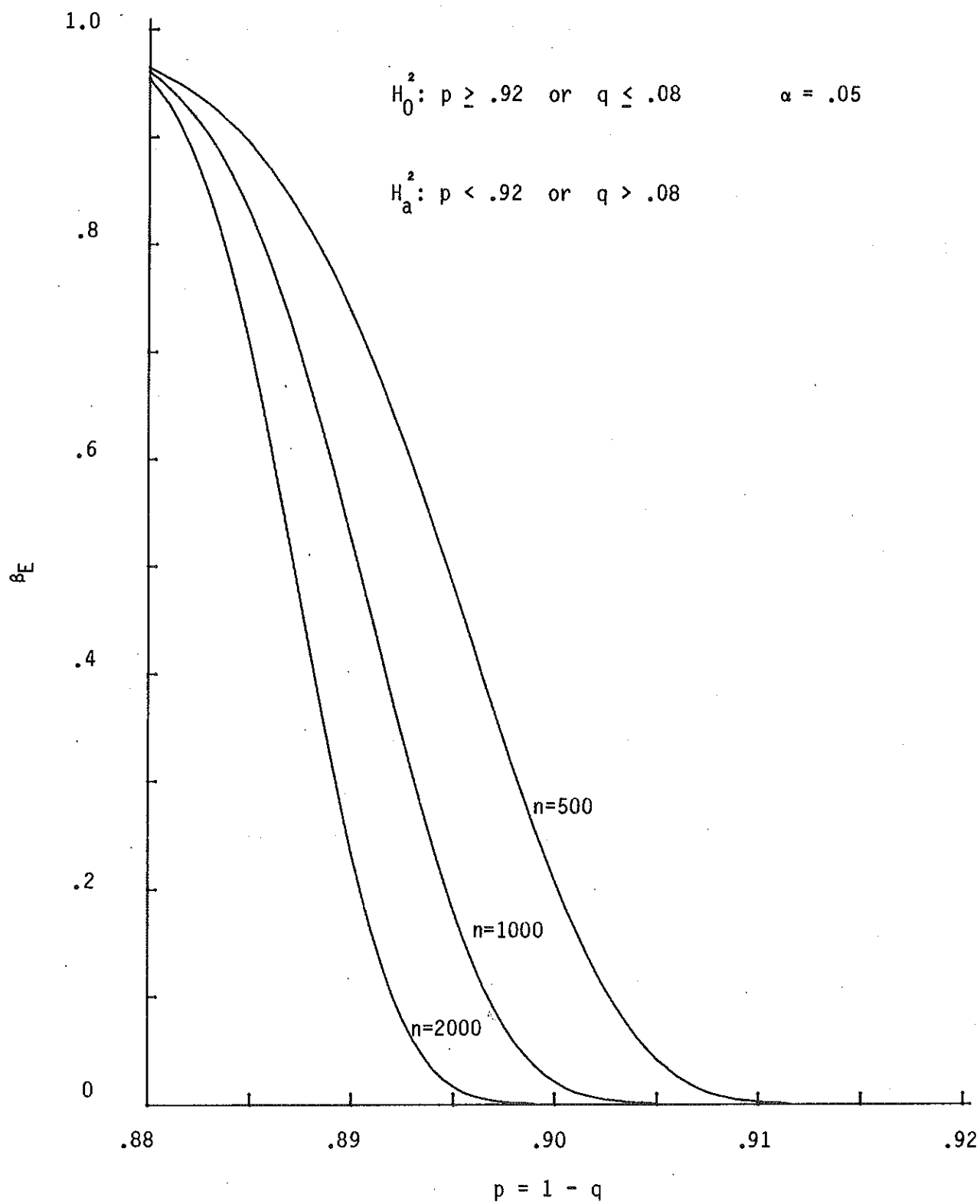


Figure 7: OC-Curve for Lower Test

Appendix II. P-VALUE AND Q-VALUE INTERPRETATION

A hypothesis test may be viewed in two distinct ways after the data has been collected. The more traditional view for performance appraisal is to merely designate above, within, or below tolerance as the evaluation for an action/task. A more informative view uses p-values⁵ and q-values⁶ to indicate the degree to which the performance is above, within, or below tolerance on one or more actions/tasks. If a job element has more than one action/task and at least one action/task is appraised using a hypothesis test, the supervisor may use p-values and q-values in the subjective mapping of action/task ratings into the job element rating. This appendix presents examples of the p-value and q-value interpretation.

If a supervisor uses a seemingly rigid hypothesis test with $p_L = .92$, $p_U = .96$, $\alpha_E = \alpha_S = .05$, $n = 2000$, $y_A = 1935$, $y_B = 1819$, $\beta_E = .0001$ for $p = .98$, and $\beta_S = .07$ for $p = .90$, the actual appraisal for this action/task can be quite flexible. Of course, the supervisor can insist that a measurement of y such that $y \geq y_A$ is needed to result in an above tolerance rating. However, a more flexible and informative interpretation might be made as follows.

Suppose that $y = 1930$ is the measurement from the sample of $n = 2000$. Since $1930 \neq 1935 = y_A$, the narrow interpretation is that the employee is not appraised as above tolerance even though $1930/2000 = .965 > .96 = p_U$.

Since the point estimate of p is greater than p_U , the employee may well be interested in how the test would have to be changed to just barely yield an above tolerance appraisal. Assuming that p_U and p_L are unchanged, both α and β risks need to be changed to make $y_A = 1930$.

The discrete nature of the binomial distribution makes it impossible to state an exact replacement for $\alpha_E = .05$. Actually, the "setting" of α_E "at" .05 really designates a range of $.046 < \alpha_E < .059$ when $n = 2000$ and $p_U = .96$. For y_A to be set at 1930, α_E must be in the range of $.138 < \alpha_E < .166$. Thus, the change needed to improve the rating requires an increase in α_E by roughly a factor of three. The formal way to make this statement is to say that (1) the p-value, as calculated from the data, is in the range of $.138 < p\text{-value} < .166$ and (2) the p-value is roughly three times the designed Type I risk,

The p-value presents one view of the data; the other view is presented by q-values. Since there are many initially designed β_E risks with each β_E corresponding to a value of p , there are many modified Type II risks when data modifies the Type I risk to a p-value. Each modified β_E risk is a q-value. All of these q-values are needed for a complete description; they may be displayed as the modified OC-curve shown in figure 1. The particular q-value of interest corresponds to $p = .965$ or $q = 1-p = .035$ because that is the point estimate provided by the measurement $y = 1930$ or $x = n-y = 70$. This q-value is .47 and corresponds to a designed β_E of .70. Thus, this q-value is roughly two-thirds of the designed Type II risk,

This particular example, and a couple of other examples which have within or above tolerance test results, are summarized in the following table:

y	Test Rating	$\frac{p\text{-value}}{q\text{-value}} \approx \frac{a_s}{(y/n)}$		$\frac{p\text{-value}}{q\text{-value}} \approx \frac{(y/n)}{p\text{-value}}$	
		$\frac{q\text{-value}}{p\text{-value}} \approx \frac{(y/n)}{a_s}$		$\frac{q\text{-value}}{p\text{-value}} \approx \frac{p\text{-value}}{(y/n)}$	
1930		$.965 \quad \frac{.138 - .166}{.046 - .059} \approx 2.3 - 3.6 \approx 3$			
	W	$\frac{.468}{.703} \approx .67 \approx 2/3$		$\frac{.468}{.138 - .166} \approx 2.8 - 3.4 \approx 3$	
1940		$.970 \quad \frac{.011 - .015}{.046 - .059} \approx .19 - .33 \approx 1/4$			
	A	$\frac{.466}{.232} \approx 2.0 \approx 2$		$\frac{.466}{.011 - .015} \approx 31 - 42 \approx 37$	
1900		$.950 \quad \frac{.988 - .991}{.9999} \approx 17 - 22 \approx 20$		$\frac{.988 - .991}{.473} \approx 2.09 - 2.10 \approx 2.1$	
	W	$\frac{.473}{.046 - .059} \approx .47 \approx 1/2$			

In the above table, the greater than unity entries for the ratio of q-value to p-value supports a final decision that the performance is above tolerance. Conversely, the greater that unity entry of p-value to q-value supports a decision of within tolerance. The magnitude of these ratios indicates the strength of this support.

The following table shows examples which have within or below tolerance ratings from the strict interpretation of hypotheses tests. A final rating of within tolerance is supported by a p-value/q-value ratio greater than unity. Conversely, q-value/p-value ratios greater than unity support a final rating of below tolerance:

	$\frac{y}{n}$	$\frac{p\text{-value}}{\alpha_E}$	$\frac{p\text{-value}}{q\text{-value} (y/n)} > 1$
y			or
Test Rating		$\frac{q\text{-value} (y/n)}{\beta_S (y/n)}$	$\frac{q\text{-value} (y/n)}{p\text{-value}} > 1$
.909		$\frac{.040 - .047}{.047 - .056} \approx .71 - 1.0 \approx 6/7$	
1818			
B		$\frac{.484}{.458} \approx 1.1 \approx 10/9$	$\frac{.489}{.040 - .047} \approx 10.4 - 12.2 \approx 11$
-	-	-	-
.915		$\frac{.215 - .240}{.047 - .056} \approx 3.8 - 5.1 \approx 9/2$	
1830			
W		$\frac{.488}{.801} \approx .61 \approx 3/5$	$\frac{.488}{.215 - .204} \approx 2.3 - 2.4 \approx 2.3$
-	-	-	-
.950		$\frac{.99999995 - .99999997}{.047 - .056} \approx (18 - 21) \approx 19$	$\frac{.99999995 - .99999997}{.486} \approx 2.1$
1900			
W		$\frac{.486}{1.000} \approx .49 \approx 1/2$	

The interpretation of the last column in both of the above tables is that the within tolerance rating is supported by a ratio of p-value/q-value that is greater than unity. This results from taking the within tolerance state as the null hypothesis. To support rejecting the null hypothesis and rate the performance as either above or below tolerance, the q-value to p-value must be greater than unity.

The bottom row in both tables is for the same measurement. This value of y , 1900, is closely within the $y_A + 1$ to $y_B - 1$ range, 1936 to 1818, which indicates neither above or below tolerance. Both p-value to q-value ratios are greater than unity and support a final rating of within tolerance. The fact that p-value/q-value ratios are essentially equal for the two tables might be unexpected since 1900 is further from $y_B = 1819$ than $y_A = 1935$. This is a consequence of having both p_L and p_U near unity; the binomial distribution is not symmetrical.

Each row in the above tables may be used to appraise performance on an individual task/action. Combinations of rows may be used in the subjective appraisal of a job element which contains several tasks/actions. Naturally, this subjective appraisal must include all tasks/actions in the job element whether or not they are treated with a hypothesis test.

As elementary examples of appraising a job element as exceeded, met, or not met, consider a job element which has only two tasks/actions. Assume that both are treated with hypothesis tests. If the two p-value to q-value ratios

are those in the $y = 1940$ and $y = 1900$ lines of the above tables, the supervisor may well subjectively decide on an exceeded rating. On the other hand, there would be less support of an exceeded rating if the ratio were from the $y = 1930$ and $y = 1900$ lines or the $y = 1940$ and $y = 1830$ lines.

Clearly, the supervisor's subjective decision becomes more complicated as the number of tasks/actions is increased. For example, a job element may have (1) a couple of tasks/actions not treated with hypothesis tests but judged within tolerance and (2) three tasks/actions with p-value to q-value ratios corresponding to those in lines of $y = 1930$, $y = 1900$, and $y = 1830$. This example has fairly strong justification for a met rating. On the other hand, replacing the $y = 1930$ line with the $y = 1818$ line would make a met appraisal more difficult to support.

In any nontrivial situation, the use of a hypothesis test on one or more task/action will not provide the supervisor with an automatic decision. The use of p-values and q-values will, however, guide the supervisor in the necessary subjective decision. Ignoring the p-values and q-values would be indefensible because that would deprive the manager of objective information.

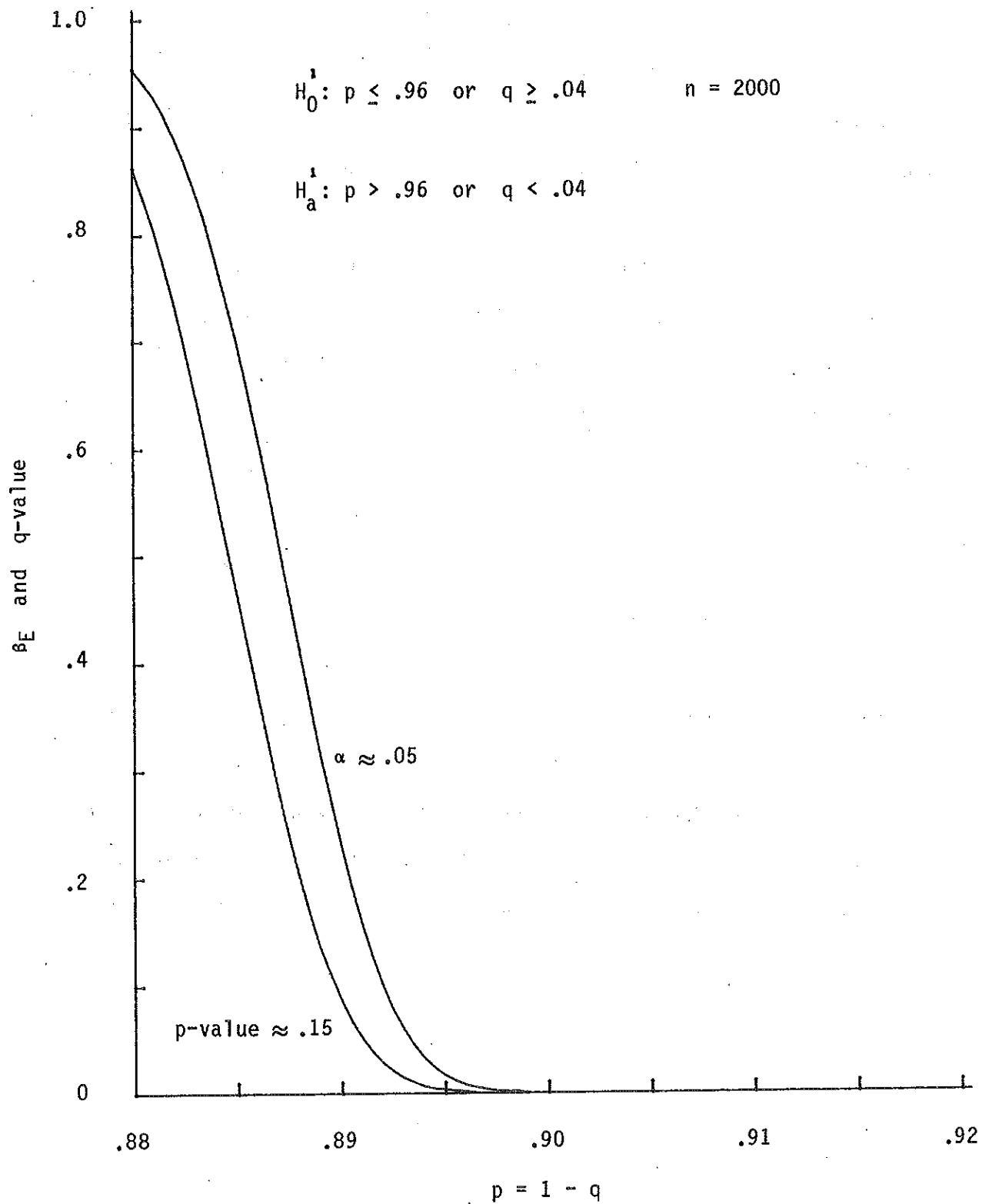


Figure 1: Q-Values for $y = 2000 - x = 1930$

MODELS FOR CONTINGENCY TABLE DATA

R.A. KOLB
DEPARTMENT OF MATHEMATICS
UNITED STATES MILITARY ACADEMY
WEST POINT, NY 10996-1786

ABSTRACT

A contingency table is a presentation of count data resulting from cross-classifications. For this type of data there are many models available to aid in the explanation of the relationships of the corresponding variables. The choice of an appropriate or, perhaps, the most appropriate model depends on a number of factors including both the generating sampling model and the hypotheses to be considered. The purpose of this paper is to describe some of these explanatory models and provide some recommendations for their use.

INTRODUCTION

The cross-classifications of a contingency table are variables, factors, or responses which have a number of levels or categories. Terms used synonymously for this type of data are cross-classified, cross-tabulated, categorical, qualitative, or frequency data. These data are the result of cross-classifying a population, or sample from a population, and accumulating totals for each "cell" of the contingency table. A cell total, then, is the number of observations from the population or sample that fall into the categorical combination represented by that cell. The table summarizes information for the entire population or sample, where every observation is categorized into one and only one cell.

A two-dimensional (two-way), $r \times s$ contingency table has two variables: one variable having r categories and one variable having s categories. The "complete" cross-classification gives a total of $r \cdot s$ cells. The following notation for a two-way, $r \times s$ table will be used:

$\{x_{ij}\} \equiv$ table of observed values;

$\{p_{ij}\} \equiv$ table of cell probabilities;

$\{m_{ij}\} \equiv$ table of expected values;

$\sum_{j=1}^s x_{ij} = x_{i.} \equiv$ observed row marginals, $i=1,2,\dots,r$;

$\sum_{i=1}^r x_{ij} = x_{.j} \equiv$ observed column marginals, $j=1,2,\dots,s$;

$\sum_{i=1}^r \sum_{j=1}^s x_{ij} = x_{..} = N \equiv$ total sample size or population.

The marginal probabilities ($p_{i.}, p_{.j}$) and marginal expected values ($m_{i.}, m_{.j}$) are similarly defined. This notation is easily extended to higher-way tables (tables with more than two variables) simply by adding more subscripts.

The primary purpose in developing models for contingency table data is to help in the determination, interpretation, and explanation of the relationships among the variables. Beginning with Pearson (1900), statistical techniques have been developed and used to test for these variable relationships, but only recently has the focus been on the use of models. Statistical techniques in support of models have now been well-developed. Specialized statistical computer packages for contingency table models (e.g. ECTA-Goodman and Fay 1973, CONTAB-Zahn 1976, and GENCAT-Landis et. al. 1976) have been available for some time and the currently popular general statistical packages (SPSS, BMDP, SAS) have contingency table data models and associated statistical techniques.

The use of these models and computer packages provides flexibility in the analysis of various type problems including those with many variables and complicated structures that a few years ago would have been impossible to analyze. The models provide the same ease of interpretation that the linear models of ANOVA and regression provide. In fact, the interpretations of the parameters of the contingency table models are often analogous to corresponding parameters in ANOVA and regression models. Also, contingency table models allow for classic model building in a manner similar to stepwise regression.

MODELS

The models available for contingency table data are many and varied and often have specialized use. The models having most universal appeal and to be discussed in this paper are the log-linear and logit models. Other models include an additive model (Bhapkar and Koch 1968), the Lancaster (1949, 1950, 1969) partitioning model, and a general linear model (Nelder and Wedderburn 1972 and Nelder 1974) with the log-linear model as a special case. The additive model has been used for special problems such as sample surveys, drug comparisons, and biological assays (e.g., see Johnson and Koch 1970 and Koch and Reinfurt 1971). Johnson and Koch discuss the advantages of the additive model for sample survey data. In general, the log-linear and logit models are the most extensively used, providing convenient parameters for most hypothesis testing situations. An excellent discussion and comparison of the corresponding additive and multiplicative interaction terms for the additive and log-linear models, respectively, is given by Darroch (1974).

The log-linear model is most convenient for general independence-type hypothesis testing situations under poisson or multinomial sampling. As a motivating example, consider a 2×2 contingency table. The classic concept of independence requires that

$$p_{ij} = p_{i.} p_{.j} \quad , \quad i = 1, 2 \quad , \quad j = 1, 2 \quad .$$

A single parameter measuring this interaction is Yule's (1900) cross-product ratio

$$\alpha = \frac{p_{11} p_{22}}{p_{12} p_{21}} \quad .$$

Independence exists when this ratio is equal to one. Taking the logarithm of α under independence,

$$\ln \alpha = \ln p_{11} - \ln p_{12} - \ln p_{21} + \ln p_{22} = 0, \quad (1)$$

we can see the motivation in using a log-linear model - a zero-valued parameter would imply independence.

The general log-linear model most frequently used was presented by Birch (1963). For an $r \times s$ table the model is

$$\ln p_{ij} = \ln p_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij); \quad i=1,2,\dots,r; \quad j=1,2,\dots,s \quad . \quad (2)$$

This model is over-parameterized in that there are $r + s + (r \cdot s) + 1$ parameters for $r \cdot s$ cells. Analogous to ANOVA, the constraints

$$\sum_i u_1(i) = \sum_j u_2(j) = \sum_i u_{12}(ij) = \sum_j u_{12}(ij) = 0 \quad (3)$$

are conveniently imposed. As an example, for the 2×2 table the constraints allow a reparametrization of the model in equation (2) by letting $u_1 = u_1(1)$, $u_2 = u_2(1)$, and $u_{12} = u_{12}(11)$, leading to the model

$$\begin{aligned} \ln p_{11} &= u + u_1 + u_2 + u_{12} \\ \ln p_{12} &= u + u_1 - u_2 - u_{12} \\ \ln p_{21} &= u - u_1 + u_2 - u_{12} \\ \ln p_{22} &= u - u_1 - u_2 + u_{12} \quad . \end{aligned} \quad (4)$$

Now the "u" parameters can be determined uniquely in terms of the logarithms of the probabilities. Specifically,

$$\begin{aligned}
 u &= 1/4 (\ell_{11} + \ell_{12} + \ell_{21} + \ell_{22}) \\
 u_1 &= 1/4 (\ell_{11} + \ell_{12} - \ell_{21} - \ell_{22}) \\
 u_2 &= 1/4 (\ell_{11} - \ell_{12} + \ell_{21} - \ell_{22}) \\
 u_{12} &= 1/4 (\ell_{11} - \ell_{12} - \ell_{21} + \ell_{22}) .
 \end{aligned} \tag{5}$$

The "u" parameters of equations (2) through (5) have analagous interpretations to the parameters of the linear model for ANOVA. In particular, for the 2×2 model of equations (4) and (5), u is the average of the logarithms of the probabilities, u_1 is the average differences across the first variable levels, and u_2 is the average differences across the second variable levels. As in ANOVA, u_{12} is an interaction term, which for the 2×2 table measures the dependence between the variables in the sense of Yules' cross-product ratio α and, specifically, from equation (1) equals $1/4 \ln \alpha$. Most importantly, under independence or "no interaction", u_{12} equals zero.

Another useful form of the log-linear model and one frequently overlooked in the literature was first presented by Ku, Varner, and Kullback (1968) and has been used primarily by Kullback and his associates. Instead of the constraints in (3), Kullback fixes one cell of the contingency table and defines the parameters to measure for each variable and interaction, a difference from this fixed cell. For the 2×2 table with cell 22 fixed, the model is

$$\begin{aligned}
 \ell_{11} &= \tau_0 + \tau_1 + \tau_2 + \tau_{12} \\
 \ell_{12} &= \tau_0 + \tau_1 \\
 \ell_{21} &= \tau_0 + \tau_2 \\
 \ell_{22} &= \tau_0 .
 \end{aligned} \tag{6}$$

Solving for the new τ parameters,

$$\begin{aligned}\tau_0 &= \ell_{22} \\ \tau_1 &= \ell_{12} - \ell_{22} \\ \tau_2 &= \ell_{21} - \ell_{22} \\ \tau_{12} &= \ell_{11} - \ell_{12} - \ell_{21} + \ell_{22}.\end{aligned}\tag{7}$$

In terms of the Birch model "u" parameters,

$$\begin{aligned}\tau_0 &= u - u_1 - u_2 + u_{12} \\ \tau_1 &= 2(u_1 - u_{12}) \\ \tau_2 &= 2(u_2 - u_{12}) \\ \tau_{12} &= 4u_{12}.\end{aligned}\tag{8}$$

The important interaction parameter τ_{12} is proportional to Birch's u_{12} and both reflect independence for values of zero.

It is interesting to recognize the similarity between these models and models for ANOVA. Similar to Birch's log linear model, the usual linear model for ANOVA defines an overall mean parameter, and measures factor effects as differences from this mean. On the other hand, similar to Kullbacks log-linear model, the regression model for ANOVA fixes one factor level, and defines the regression coefficients as the differences of the other factors from this fixed level.

In addition to log-linear models, logit models are also very popular for certain applications of contingency tables. In particular, for product-multinomial sampling with homogeneity-type hypotheses and one or more response variables, logit models are very useful. For example, consider the factor and response problem depicted in Figure 1.

		B		
		1	2	
A	1	p_{11}	p_{12}	1
	2	p_{21}	p_{22}	1

Figure 1, Factor A, Response B

Here, A is the factor at 2 levels and B is the binomial response. The homogeneity hypothesis would be $H_0: p_{11} = p_{21}$ (or $p_{12} = p_{22}$). Under H_0 , the log-linear models would require that two parameters equal zero, namely from (5) and (7),

$$\text{Birch: } u_1 = u_{12} = 0$$

$$\text{Kullback: } \tau_1 = \tau_{12} = 0.$$

Yet, the homogeneity hypothesis is a one degree-of-freedom test and a convenient model should provide a single corresponding zero-valued parameter. Defining the logit $L_i = \ln(p_{i1}/p_{i2})$ for $i = 1, 2$,

$$\begin{aligned} L_1 &= \ln(p_{11}/p_{12}) = \ln p_{11} - \ln p_{12} \\ &= 2u_2 + 2u_{12} \end{aligned}$$

and

$$\begin{aligned} L_2 &= \ln(p_{21}/p_{22}) = \ln p_{21} - \ln p_{22} \\ &= 2u_2 - 2u_{12}. \end{aligned}$$

Letting $w = 2u_2$ and $w_1 = 2u_{12}$,

$$L_1 = w + w_1 \tag{9}$$

$$L_2 = w - w_1$$

and

$$w = 1/2 (L_1 + L_2)$$

$$w_1 = 1/2 (L_1 - L_2). \tag{10}$$

Now, the single model parameter w_1 corresponds to the one degree-of-freedom homogeneity hypothesis (i.e., $H_0: p_{11} = p_{21} \Leftrightarrow w_1 = 0$).

LARGER TABLES

Extending these models to larger tables is relatively straight forward; although, some care is required to insure clear definitions of the parameters so that they will purposely relate to the hypotheses of concern. Appendix A provides the models and hypotheses for the 2×3 table and Appendix B for the three-way $2 \times 2 \times 2$ table.

Initially, considering the 2×3 table, the independence hypothesis is a two degree of freedom test and each log-linear model provides two convenient parameters, u_{12} and u'_{12} for the Birch model and τ_{11}^{ij} and τ_{12}^{ij} for the Kullback model. In comparing the models, the arbitrary fixing of a cell in the Kullback model may not appeal to some analysts, but the relative simplicity of the model would certainly appeal to all. The independence parameters for the Kullback model are also easier to interpret. Letting α_{mn} be the cross product ratio of column m and column n taken as a 2×2 table, independence occurs when the three cross product ratios α_{12} , α_{13} , and α_{23} are equal to one (any two α_{mn} equal to one will insure that the third is equal to one). The log-linear parameters relate to these α_{mn} in the following manner:

$$u_{12} = 1/6 (\ln \alpha_{12} + \ln \alpha_{13})$$

$$u'_{12} = 1/6 (\ln \alpha_{12} + \ln \alpha_{23})$$

$$\tau_{11}^{ij} = \ln \alpha_{13}$$

$$\tau_{12}^{ij} = \ln \alpha_{23} .$$

The Kullback τ parameters are simply the logarithms of Yules' original cross-product ratios for the 2×2 subtables that include the fixed cell.

The appropriate logit model is dependent on the scheme of sampling. When the data is sampled across the rows, it is convenient to build a model that calculates logits based on ratios of row probabilities for each column. This is reflected in the III.a. model of Appendix A. Symmetrically, when data is sampled across columns, it is convenient to build a model that calculates logits based on ratios of column probabilities for each row. This is reflected in the III.b. model of Appendix A. For the sampling model in III.a., the corresponding homogeneity hypothesis is a two degree of freedom test that compares the probabilities across a row. The logit model provides the three parameters w_1 , w_2 , and w_3 and the constraint that their sum equals zero. For the model in III.b., the homogeneity hypothesis is a two-degree of freedom test that compares the probabilities across any two of the three columns. The logit model provides three parameters (corresponding to the three columns); any two of which can be used to test the hypothesis. It should be noted that other logit parameterizations are possible.

Turning now to the three-way $2 \times 2 \times 2$ table in Appendix B, the comparative simplicity of the Kullback model is again apparent. In the Kullback model the 222 cell has been fixed. The main effects $(\tau_1^i, \tau_1^j, \tau_1^k)$ measure the difference between the second and first levels of each variable as compared to the fixed cell. The two-way interaction terms $(\tau_{11}^{ij}, \tau_{11}^{ik}, \tau_{11}^{jk})$ are the logarithms of the three possible cross-product ratios with the 222 cell that measure interaction between two variables with the third fixed. The three-way interaction term (τ_{111}^{ijk}) is the difference of the logarithms of the cross-product ratios when variable one is fixed at level one compared to level two.

The Birch model uses a mean parameter (u) which is the average of the logarithms of the cell probabilities. The main effects (u_1, u_2, u_3) average the difference in the logarithms of the probabilities at the two levels of each variable, respectively. The interaction terms (u_{12}, u_{13}, u_{23}) average the logarithms of the cross-product ratios corresponding to the two measured variables. The three-way interaction term (u_{123}) measures the same difference of logarithms of cross-product ratios as does τ_{111}^{ijk} ; although, it averages this difference across the cells by taking 1/8 the value.

The presented logit model considers that variable one is a response variable and that product-multinomial sampling is appropriate. The model is analogous to the 2×2 Birch log-linear model; however, the parameters (w_2, w_3, w_{23}) measure the effect that the corresponding terms have on the response variable.

Considering the hypotheses for the $2 \times 2 \times 2$ table as listed in paragraph IV of Appendix B, the no three-way interaction hypothesis is a one degree-of-freedom test and each model provides one corresponding parameter. The logit model w_{23} parameter (and corresponding hypothesis test) is more properly interpreted as a measure of the interaction between variables two and three as it affects variable one. The mutual independence test under multinomial sampling is a four degree of freedom test and the two log-linear models provide four parameters corresponding to each possible interaction. Under product-multinomial sampling the test has three degrees of freedom and the logit model provides three parameters. The conditional independence test requires that one variable be considered fixed and that independence between the other two-variables be tested. In Appendix B, variable three has been fixed. This is a two-degree of freedom test and each model provides two parameters. The homogeneity test

has many forms. The one chosen in Appendix B corresponds to the selection of variable one as the response variable in the logit model. Under complete homogeneity, all these logits and logit parameters are equal to zero. In effect the $2 \times 2 \times 2$ table has collapsed to a 2×2 table with variables two and three remaining. The terms of the log-linear models relating to the first variable are also now zero.

CONCLUSION

It might be said that there is only a limited amount of information available from any given data set. For contingency table data, the models presented in this paper provide the means to fully explain the data with respect to the measured variables, and often indicate relationships which might not have been apparent with other techniques.

REFERENCES

- Bhapkar, V.P., and G.G. Koch (1968), "On the Hypotheses of 'No Interaction' in Contingency Tables", Biometrics, 24, 567-594.
- Birch, M.W. (1963), "Maximum Likelihood in Three-Way Contingency Tables", Journal of the Royal Statistical Society (B), 25, 220-223.
- Darroch, J.N. (1974), "Multiplicative and Additive Interaction in Contingency Tables", Biometrika, 61, 207-214.
- Goodman, L.A., and R. Fay (1973), Everyman's Contingency Table Analysis: Program Documentation, Chicago: The University of Chicago Press.
- Johnson, W.D., and G.G. Koch (1970), "Analysis of Qualitative Data: Linear Functions", Health Sciences Research, 5, 358-369.
- Koch, G.G., and D.W. Reinfurt (1971), "The Analysis of Complex Contingency Table Data from General Experimental Designs and Sample Surveys", Proceedings of the 16th Conference on the Design of Experiments in Army Research Development and Testing, ARO-D Report 71-3, 453-527.
- Ku, H.H., R.H. Varner, and S. Kullback (1971), "On the Analysis of Multidimensional Contingency Tables", Journal of the American Statistical Association, 66, 55-64.

- Kullback, S. (1959), Information Theory and Statistics, New York: John Wiley and Sons, Inc.
- Lancaster, H.O. (1949), "The Derivation and Partition of X^2 in Certain Discrete Distributions", Biometrika, 36, 117-129.
- Lancaster, H.O. (1950), "The Exact Partition of X^2 and It's Application to the Problem of the Pooling of Small Expectations", Biometrika, 37, 267-270.
- Lancaster, H.O. (1969), "Contingency Tables of Higher Dimensions", Bulletin of the International Statistical Institute, 43, 143-151.
- Landis, J.R., W.M. Stanish, J.L. Freeman, and G.G. Koch (1976), "A Computer Program for the Generalized Chi-Square Analysis of Categorical Data Using Weighted Least Squares (GENCAT)", Computer Programs in Biomedicine, 6, 196-231.
- Nelder, J.A. (1974), "Log-Linear Models for Contingency Tables: A Generalization of Classical Least Squares", Applied Statistics, 23, 323-329.
- Nelder, J.A., and R.W.M. Wedderburn (1972), "Generalized Linear Models", Journal of the Royal Statistical Society, 135, 370-384.
- Pearson, K. (1900), "On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling", Philosophical Magazine, Series 5, 50, 157-175.
- Yule, G.U. (1900), "On the Association of Attributes in Statistics: With Illustration from the Material of the Childhood Society", Philosophical Transactions of the Royal Society of London (A), 194, 257-319.
- Zahn, D.A. (1976), Documentation for CONTAB-A Computer Program to Aid in the Analysis of Multi-Dimensional Contingency Tables Using Log-Linear Models, FSU Statistics Report M292, Department of Statistics and Statistical Consulting Center, The Florida State University, Tallahassee, Florida.

APPENDIX A. 2 x 3 TABLE MODELS

p_{11}	p_{12}	p_{13}
p_{21}	p_{22}	p_{23}

I. Birch Log-linear Model

General:

$$\ell_{ij} = \ln p_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij); i = 1, 2; j = 1, 2, 3.$$

Define:

$$u_1 = u_1(1) \quad u_{12} = u_{12}(11)$$

$$u_2 = u_2(1) \quad u'_{12} = u_{12}(12)$$

$$u'_2 = u_2(2)$$

Model:

$$\ell_{11} = u + u_1 + u_2 + u_{12}$$

$$\ell_{12} = u + u_1 + u'_2 + u'_{12}$$

$$\ell_{13} = u + u_1 + u_2 - u'_2 - u_{12} - u'_{12}$$

$$\ell_{21} = u - u_1 + u_2 - u_{12}$$

$$\ell_{22} = u - u_1 + u'_2 - u'_{12}$$

$$\ell_{23} = u - u_1 - u_2 - u'_2 + u_{12} + u'_{12}$$

Parameters:

$$u = 1/6 (\ell_{11} + \ell_{12} + \ell_{13} + \ell_{21} + \ell_{22} + \ell_{23})$$

$$u_1 = 1/6 (\ell_{11} + \ell_{12} + \ell_{13} - \ell_{21} - \ell_{22} - \ell_{23})$$

$$u_2 = 1/6 (2\ell_{11} - \ell_{12} - \ell_{13} + 2\ell_{21} - \ell_{22} - \ell_{23})$$

$$u'_2 = 1/6 (-\ell_{11} + 2\ell_{12} - \ell_{13} - \ell_{21} + 2\ell_{22} - \ell_{23})$$

$$u_{12} = 1/6 (2\ell_{11} - \ell_{12} - \ell_{13} - 2\ell_{21} + \ell_{22} + \ell_{23})$$

$$u'_{12} = 1/6 (-\ell_{11} + 2\ell_{12} - \ell_{13} + \ell_{21} - 2\ell_{22} + \ell_{23})$$

II. Kullback Log-linear Model

Define: Cell 23 fixed

$$\text{Model: } \ell_{11} = \tau_0 + \tau_1^1 + \tau_1^j + \tau_{11}^{1j}$$

$$\ell_{12} = \tau_0 + \tau_1^1 + \tau_2^j + \tau_{12}^{1j}$$

$$\ell_{13} = \tau_0 + \tau_1^1$$

$$\ell_{21} = \tau_0 + \tau_1^j$$

$$\ell_{22} = \tau_0 + \tau_2^j$$

$$\ell_{23} = \tau_0$$

Parameters:

$$\tau_0 = \ell_{23}$$

$$\tau_1^1 = \ell_{13} - \ell_{23}$$

$$\tau_1^j = \ell_{21} - \ell_{23}$$

$$\tau_2^j = \ell_{22} - \ell_{23}$$

$$\tau_{11}^{1j} = \ell_{11} - \ell_{13} - \ell_{21} + \ell_{23}$$

$$\tau_{12}^{1j} = \ell_{12} - \ell_{13} - \ell_{22} + \ell_{23}$$

III. Logit Model

$$\text{a. } \sum_{i=1}^2 p_{ij} = 1 \text{ for } j = 1, 2, 3$$

$$\text{Define: } L_j = \ln(p_{1j}/p_{2j}), j = 1, 2, 3$$

$$\text{Model: } L_1 = w + w_1$$

$$L_2 = w + w_2$$

$$L_3 = w + w_3$$

Constraint:

$$w_1 + w_2 + w_3 = 0$$

Parameters:

$$w = 1/3 (L_1 + L_2 + L_3)$$

$$w_1 = 1/3 (2L_1 - L_2 - L_3)$$

$$w_2 = 1/3 (-L_1 + 2L_2 - L_3)$$

$$w_3 = 1/3 (-L_1 - L_2 + 2L_3)$$

b. $\sum_{j=1}^3 p_{ij} = 1$ for $i = 1, 2$

Define: $L_{ij} = \ln p_{ij} / \sum_{k \neq j} p_{ik}$ for $i = 1, 2; j = 1, 2, 3$.

General: $L_{ij} = w + w_j(i)$

Constraints: $\sum_{i=1}^2 w_j(i) = 0$ for $j = 1, 2, 3$

Define: $w_1 = w_1(1), w_2 = w_2(1), w_3 = w_3(1)$

Model: $L_{11} = w + w_1$

$$L_{12} = w + w_2$$

$$L_{13} = w + w_3$$

$$L_{21} = w - w_1$$

$$L_{22} = w - w_2$$

$$L_{23} = w - w_3$$

Parameters: $w = 1/6 \sum_{ij} L_{ij}$

$$w_1 = 1/2 (L_{11} - L_{21})$$

$$w_2 = 1/2 (L_{12} - L_{22})$$

$$w_3 = 1/2 (L_{13} - L_{23})$$

IV. Hypotheses

1. Independence H_0 : $p_{ij} = p_{i.} p_{.j}$, $i = 1, 2$; $j = 1, 2, 3$

Birch H_0 : $u_{12} = u'_{12} = 0$

Kullback H_0 : $\tau_{11}^{ij} = \tau_{12}^{ij} = 0$

2. Homogeneity

a. H_0 : $p_{11} = p_{12} = p_{13} \Rightarrow p_{21} = p_{22} = p_{23}$

Birch H_0 : $u_2 = u'_2 = u_{12} = u'_{12} = 0$

Kullback H_0 : $\tau_1^j = \tau_2^j = \tau_{11}^{ij} = \tau_{12}^{ij} = 0$

Logit H_0 : $w_1 = w_2 = w_3 = 0$

b. H_0 : $p_{11} = p_{21} \Rightarrow p_{13} = p_{23}$
 $p_{12} = p_{22}$

Birch H_0 : $u_1 = u_{12} = u'_{12} = 0$

Kullback H_0 : $\tau_1^i = \tau_{11}^{ij} = \tau_{12}^{ij} = 0$

Logit H_0 : $w_1 = w_2 = w_3 = 0$

APPENDIX B. $2 \times 2 \times 2$ TABLE MODELS

p_{111}	p_{112}
p_{121}	p_{122}

p_{211}	p_{212}
p_{221}	p_{222}

I. Birch Log-linear Model

General:

$$l_{ijk} = u + u_1(i) + u_2(j) + u_3(k) + u_{12}(ij) + u_{13}(ik) + u_{23}(jk) + u_{123}(ijk);$$

$$i = 1, 2; j = 1, 2; k = 1, 2.$$

Define:

$$\begin{aligned} u_1 &= u_1(1) & u_3 &= u_3(1) & u_{12} &= u_{12}(11) \\ u_2 &= u_2(1) & u_{12} &= u_{12}(11) & u_{23} &= u_{23}(11) \end{aligned}$$

Model:

$$\begin{aligned} l_{111} &= u + u_1 + u_2 + u_3 + u_{12} + u_{13} + u_{23} + u_{123} \\ l_{112} &= u + u_1 + u_2 - u_3 + u_{12} - u_{13} - u_{23} - u_{123} \\ l_{121} &= u + u_1 - u_2 - u_3 - u_{12} + u_{13} - u_{23} - u_{123} \\ l_{211} &= u - u_1 + u_2 + u_3 - u_{12} - u_{13} + u_{23} - u_{123} \\ l_{212} &= u - u_1 + u_2 - u_3 - u_{12} + u_{13} - u_{23} + u_{123} \\ l_{221} &= u - u_1 - u_2 + u_3 + u_{12} - u_{13} - u_{23} + u_{123} \\ l_{222} &= u - u_1 - u_2 - u_3 + u_{12} + u_{13} + u_{23} - u_{123} \end{aligned}$$

Parameters:

$$\begin{aligned} u &= 1/8 (l_{111} + l_{112} + l_{121} + l_{122} + l_{211} + l_{212} + l_{221} + l_{222}) \\ u_1 &= 1/8 (l_{111} + l_{112} + l_{121} + l_{122} - l_{211} - l_{212} - l_{221} - l_{222}) \\ u_2 &= 1/8 (l_{111} + l_{112} - l_{121} - l_{122} + l_{211} + l_{212} - l_{221} - l_{222}) \\ u_3 &= 1/8 (l_{111} - l_{112} + l_{121} - l_{122} + l_{211} - l_{212} + l_{221} - l_{222}) \\ u_{12} &= 1/8 (l_{111} + l_{112} - l_{121} - l_{122} - l_{211} - l_{212} + l_{221} + l_{222}) \\ u_{13} &= 1/8 (l_{111} - l_{112} + l_{121} - l_{122} - l_{211} + l_{212} - l_{221} + l_{222}) \\ u_{23} &= 1/8 (l_{111} - l_{112} - l_{121} + l_{122} + l_{211} - l_{212} - l_{221} + l_{222}) \\ u_{123} &= 1/8 (l_{111} - l_{112} - l_{121} + l_{122} - l_{211} + l_{212} + l_{221} - l_{222}) \end{aligned}$$

II. Kullback log-linear Model.

Define: Cell 222 fixed

Model:

$$\ell_{111} = \tau_0 + \tau_1^i + \tau_1^j + \tau_1^k + \tau_{11}^{ij} + \tau_{11}^{ik} + \tau_{11}^{jk} + \tau_{111}^{ijk}$$

$$\ell_{112} = \tau_0 + \tau_1^i + \tau_1^j + \tau_{11}^{ij}$$

$$\ell_{121} = \tau_0 + \tau_1^i + \tau_1^k + \tau_{11}^{ik}$$

$$\ell_{122} = \tau_0 + \tau_1^i$$

$$\ell_{211} = \tau_0 + \tau_1^j + \tau_1^k + \tau_{11}^{jk}$$

$$\ell_{212} = \tau_0 + \tau_1^j$$

$$\ell_{222} = \tau_0$$

Parameters:

$$\tau_0 = \ell_{222}$$

$$\tau_1^i = \ell_{122} - \ell_{222}$$

$$\tau_1^j = \ell_{212} - \ell_{222}$$

$$\tau_1^k = \ell_{221} - \ell_{222}$$

$$\tau_{11}^{ij} = \ell_{112} - \ell_{122} - \ell_{212} + \ell_{222}$$

$$\tau_{11}^{ik} = \ell_{121} - \ell_{122} - \ell_{221} + \ell_{222}$$

$$\tau_{11}^{jk} = \ell_{211} - \ell_{212} - \ell_{221} + \ell_{222}$$

$$\tau_{111}^{ijk} = \ell_{111} - \ell_{112} - \ell_{121} + \ell_{122} - \ell_{211} + \ell_{212} + \ell_{221} - \ell_{222}$$

III. Logit Model:

General:

$$L_{jk} = \ln\left(\frac{p_{1jk}}{p_{2jk}}\right) = w + w_2(j) + w_3(k) + w_{23}(jk); j = 1, 1; k = 1, 2$$

where

$$\sum_j w_2(j) = \sum_k w_3(k) = \sum_j w_{23}(jk) = \sum_k w_{23}(jk) = 0.$$

Define:

$$w_2 = w_2(1), w_3 = w_3(1), w_{23} = w_{23}(11).$$

Model:

$$L_{11} = w + w_2 + w_3 + w_{23}$$

$$L_{12} = w + w_2 - w_3 - w_{23}$$

$$L_{21} = w - w_2 + w_3 - w_{23}$$

$$L_{22} = w - w_2 - w_3 + w_{23}$$

Parameters:

$$w = 1/4 (L_{11} + L_{12} + L_{21} + L_{22})$$

$$w_2 = 1/4 (L_{11} + L_{12} - L_{21} - L_{22})$$

$$w_3 = 1/4 (L_{11} - L_{12} + L_{21} - L_{22})$$

$$w_{23} = 1/4 (L_{11} - L_{12} - L_{21} + L_{22})$$

IV. Hypotheses

1. No three-way Interaction (No second order Interaction)

$$\text{Birch } H_0: u_{123} = 0$$

$$\text{Kullback } H_0: \tau_{111}^{ijk} = 0$$

$$\text{Logit } H_0: w_{23} = 0$$

2. Mutual (Complete) Interaction

$$\text{Birch } H_0: u_{12} = u_{13} = u_{23} = u_{123} = 0$$

$$\text{Kullback } H_0: \tau_{11}^{ij} = \tau_{11}^{ik} + \tau_{11}^{jk} = \tau_{111}^{ijk} = 0$$

$$\text{Logit } H_0: w_2 = w_3 = w_{23} = 0$$

3. Conditional Independence (1 to 2 with 3 fixed)

$$\text{Birch } H_0: u_{12} = u_{123} = 0$$

$$\text{Kullback } H_0: \tau_{11}^{ij} = \tau_{111}^{ijk} = 0$$

$$\text{Logit } H_0: w_2 = w_{23} = 0$$

4. Homogeneity of Tables

$$\text{Birch } H_0: u_1 = u_{12} = u_{13} = u_{123} = 0$$

$$\text{Kullback } H_0: \tau_1^i = \tau_{11}^{ij} = \tau_{11}^{ik} = \tau_{111}^{ijk} = 0$$

$$\text{Logit } H_0: w = w_2 = w_3 = w_{23} = 0$$

On a Class of Probability Density Functions

H. P. Dudel and S. H. Lehnigk

U.S. Army Missile Command
Research, Development, and Engineering Center
Research Directorate
Redstone Arsenal, Alabama 35898-5248

SUMMARY

The application of a three parameter class of one-sided probability distributions is being discussed. For specific parameter values, this class contains as special cases a number of well-known distributions of statistics and statistical physics, namely, Gauss, Weibull, exponential, Rayleigh, Gamma, chi-square, Maxwell, and Wien (limiting case of Planck's distribution). One of the three parameters represents scale; the other two represent initial and terminal shape of the associated probability density function. A fourth parameter, shift, may be introduced. The distribution class discussed in this paper was introduced by L. Amoroso [2] in 1924. It is closely connected with a family of linear Fokker-Planck equations (generalized Feller equation). In fact, the class of probability density functions associated with the distribution class considered here is a special case of the set of all delta function initial condition solutions of the generalized Feller equation for a fixed value of the time variable. It will be shown that, as a function of the logarithm of the independent variable, the logarithm of the cumulative distribution function is asymptotically linear as the independent variable approaches zero from above. This fact leads to a general criterion for the applicability of the presented distribution family relative to given empirical data. The applicability criterion can be used to determine approximate values for the two shape parameters. They can subsequently be used as initial values in any of the established parameter estimation techniques.

1. A Class of Distributions

A number of basic continuous distributions of classical statistics and statistical physics are special cases of a class of distributions which is characterized by the cumulative distribution function

$$F(x) = \begin{cases} \frac{1}{\Gamma(1+q)} \gamma(1+q, \xi^{1-\lambda}) , & \xi = xb^{-1}, q = (\lambda-p)(1-\lambda)^{-1}, x > 0, \\ 0 , & x \leq 0, \end{cases} \quad (1.1)$$

which depends on the three mutually independent parameters $b > 0$, $p < 1$, and $\lambda < 1$. With these restrictions on the parameters p and λ , the composite quantity $q = (\lambda-p)(1-\lambda)^{-1}$ will be greater than -1 . In standard terminology, b is the scale parameter, and there are two shape parameters, λ and p , which are independent of each other. A fourth parameter, the shift parameter x_0 , may be introduced by replacing x by $x - x_0$. The functions $\Gamma(y)$ and $\gamma(a, y)$ in (1.1) are the Gamma and the incomplete Gamma functions, respectively.

By means of the integral definition of $\gamma(a, y)$ [1,8.350.1], (1.1) can be expressed in the form

$$F(x) = \frac{1}{\Gamma(1+q)} \int_0^{\xi^{1-\lambda}} t^{(1-q)-1} e^{-t} dt, \quad x > 0. \quad (1.2)$$

Since $\gamma(a, y)$ may also be defined by means of the degenerate hypergeometric function $\Phi(= F)$ [1,9.236.4, 9.210.1],

$$\gamma(a, y) = \frac{1}{a} y^a \Phi(a, a+1; -y), \quad (1.3)$$

we obtain a third expression for $F(x)$,

$$F(x) = \frac{1}{\Gamma(2+q)} \xi^{1-p} \Phi(1+q, 2+q; -\xi^{1-\lambda}), \quad x > 0, \quad (1.2)$$

which will turn out to be quite useful later on.

The probability density function $f(x)$ associated with the cumulative distribution function $F(x)$ is given by

$$f(x) = \begin{cases} \frac{1-\lambda}{\Gamma(1+q)} b^{-1} \xi^{-p} \exp -\xi^{1-\lambda}, & \xi = xb^{-1}, q = (\lambda-p)(1-\lambda)^{-1}, x > 0, \\ 0, & x \leq 0. \end{cases} \quad (1.3)$$

The distribution class defined by either the cumulative distribution function (1.1) or the probability density function (1.3) was introduced by L. Amoroso [2] in 1924 and reconsidered in later publications, [3], [4], [5], and [6].

Some other aspects of this density function class have been discussed in [7] from a theoretical point of view. That paper contains remarks about the associated probability measure space and the associated characteristic function class. A more thorough discussion of the characteristic functions from the point of view of complex function theory will be presented elsewhere [8].

The class of density functions (1.3) contains the following special cases: Gauss (normal), Weibull, exponential, Rayleigh, Gamma, chi-square, Maxwell, and Wien, as has been pointed out in [7].

2. The Moments

All moments of the distribution class characterized either by the cumulative distribution $F(x)$ given in (1.1) or by the associated density function $f(x)$ given in (1.3) exist, provided the parameters b , λ , and p are kept within the ranges $b > 0$, $\lambda < 1$, and $p < 1$.

The characteristic function associated with $F(x)$ and $f(x)$ is given by the Laplace integral

$$\begin{aligned}\Psi(s) &= \int_0^{\infty} f(x) e^{sx} dx \\ &= \frac{1-\lambda}{\Gamma(1+q)} b^{-1} \int_0^{\infty} \xi^{-p} \exp(-\xi^{1-\lambda} + sx) dx \\ &= \frac{1-\lambda}{\Gamma(1+q)} \int_0^{\infty} \xi^{-p} \exp(-\xi^{1-\lambda} + sb\xi) d\xi, \quad \xi = xb^{-1},\end{aligned}\tag{2.1}$$

where s is a complex variable. The last integral in (2.1) converges for $\operatorname{Re} s \leq 0$ if $0 < \lambda < 1$, for $\operatorname{Re} s < b^{-1}$ if $\lambda = 0$, and for every s if $\lambda < 0$. Reference is made to [7] and for a more detailed investigation, to [8]. It follows that $\Psi(s)$ is holomorphic in the domain $\operatorname{Re} s < 0$ if $0 < \lambda < 1$, in $\operatorname{Re} s < b^{-1}$ if $\lambda = 0$, and it is an entire function if $\lambda < 0$. Therefore, for $\lambda \leq 0$ the moments of our distribution class are given by

$$\begin{aligned}
m_n = \psi^{(n)}(0) &= \frac{b^n}{\Gamma(1+q)} \int_0^\infty t^{(n-p+1)(1-\lambda)^{-1}-1} e^t dt \\
&= \frac{b^n}{\Gamma(1+q)} \Gamma\left(1 + q + \frac{n}{1-\lambda}\right) \quad (n=0,1,2,\dots). \quad (2.2)
\end{aligned}$$

In particular, $m_0 = 1$, and the first moment, or mean μ , is

$$m_1 = \mu = \frac{b}{\Gamma(1+q)} \Gamma\left(1 + q + \frac{1}{1-\lambda}\right). \quad (2.3)$$

If λ is in the range $0 < \lambda < 1$, $\Psi(s)$ is not holomorphic at $s=0$. There is no power series expansion about $s=0$. The moments in this situation may still be defined, however, by (2.2) as $\lim \psi^{(n)}(s_0)$, $\text{Re } s_0 < 0$, as $s_0 \rightarrow 0$ two-dimensionally in the left-hand s -plane. Of course, one may alternatively use the definition of the moments in the form

$$m_n = \int_0^\infty x^n f(x) dx \quad (n=0,1,2,\dots)$$

for $0 < \lambda < 1$.

3. An Associated Differential Equation

From an application point of view the usefulness of the distribution function defined in Section 1 lies in the fact that it contains two independent shape parameters, p and λ , which allows fitting initial and terminal shapes (in the direction of increasing x) of given distribution data independently. However, there is another aspect which may very well be of fundamental theoretical interest.

The class of density functions (1.3) is closely connected to a class of Fokker-Planck equations. By fiat this connection then is typical for all of the special cases listed in Section 1. It makes it possible to investigate the underlying probabilistic features of the function class (1.3) and its special cases by employing the machinery of probability theory.

Disregarding statistical considerations completely at this point, one may ask the question: what is the most general one-dimensional autonomous parabolic (Fokker-Planck) equation

$$\frac{\partial}{\partial x} \left[A(x) \frac{\partial z}{\partial x} + D(x)z \right] - \frac{\partial z}{\partial t} = 0, \quad z = z(x,t), \quad x > 0, \quad t > 0, \quad (3.1)$$

which admits a similarity solution

$$z_0(x,t) = b^{-1}(t)f^*(\xi), \quad \xi = xb^{-1}(t), \quad (3.2)$$

which is conservative, i.e., for which

$$\int_0^{\infty} z_0(x,t)dx \equiv 1.$$

This question is an important one in the attempt to model diffusion processes in the applied sciences and to define initial and boundary condition solutions of an equation of the form (3.1). In practical terms, the coefficients $A(x)$ and $D(x)$ in (3.1) are the diffusion and drift coefficients, respectively. $D(x)$ is being called the drift coefficient because, if x has the unit length and t the unit time, then $D(x)$ acquires the unit length/time.

To obtain conditions for the coefficients $A(x)$ and $D(x)$ and for the functions $f^*(\xi)$ and $b(t)$ appearing in (3.2), we substitute $z_0(x,t)$ into the equation (3.1) and obtain a first order ordinary equation involving $A(x)$ and $D(x)$, a second order ordinary equation for $f^*(\xi)$, and a first order ordinary equation for $b(t)$. In the absence of any further conditions on $z_0(x,t)$, the differential relationship between $A(x)$ and $D(x)$ cannot be uniquely solved. Practical considerations in a number of specific situations required the diffusion coefficient to obey a power law of the form

$$A(x) = \alpha x^{1+\lambda}, \quad \alpha > 0. \quad (3.3)$$

The drift coefficient then becomes

$$D(x) = \alpha p x^\lambda + \beta x, \quad \lambda < 1, \quad p < 1, \quad \beta \in \mathbb{R}. \quad (3.4)$$

The resulting equation for $f^*(\xi)$ has the particular solution

$$f^*(\xi) = \frac{1-\lambda}{\Gamma(1+q)} \xi^{-p} \exp - \xi^{1-\lambda}, \quad q = (\lambda-p)(1-\lambda)^{-1}, \quad (3.5)$$

and the function $b(t)$ becomes

$$b(t) = \begin{cases} [\alpha(1-\lambda)^2 t]^{(1-\lambda)^{-1}}, & \beta = 0, \\ [\alpha(1-\lambda)\beta^{-1} (1 - \exp - (1-\lambda)\beta t)]^{(1-\lambda)^{-1}}, & \beta \neq 0. \end{cases} \quad (3.6)$$

Mathematical aspects of the differential equation (3.1) with its coefficients specified by (3.3) and (3.4), which has been designated generalized

Feller equation have been investigated in a sequence of papers [9], [10], [11], [12]. The special types of the equation (3.1), (3.3), (3.4) for the cases of interest in statistics in connection with the special distributions listed in Section 1 have been given in [7].

Within the framework of this paper it is of interest to note that

(1) The function $z_0(x, t)$ in (3.2) with $f^*(\xi)$ and $b(t)$ specified in (3.5) and (3.6), respectively, i.e.,

$$z_0(x, t) = \frac{1-\lambda}{\Gamma(1+q)} b^{-1} \xi^{-p} \exp -\xi^{1-\lambda}, \quad (3.7)$$

is the delta function initial condition solution of (3.1), (3.3), (3.4), with the delta function applied at $x=0$, $t=0$ [9]. In other words, the similarity solution (3.7) describes the distribution process governed by (3.1), (3.3), (3.4) from a completely concentrated initial state at $x=0$, $t=0$.

(2) If we "stop" this process at any time $t_0 > 0$, we see that, setting $b(t_0) = b$ and comparing (3.7) and (1.3), the function $z_0(x, t_0)$ becomes the probability density function $f(x)$ of the process at $t = t_0$. This fact opens up the intriguing opportunity of studying the statistical or probabilistic behavior of the underlying process in time if the scale parameter b is allowed to vary according to (3.6).

(3) It is easily seen from (3.6) that $b(t) \uparrow + \infty$ as $t \uparrow + \infty$ if the drift parameter $\beta \leq 0$. This means the process will "spread out" over the entire positive x -axis. However, if $\beta > 0$, $b(t) \uparrow [\alpha(1-\lambda)\beta^{-1}]^{(1-\lambda)^{-1}}$, a finite constant, as $t \uparrow + \infty$. In other words, the process approaches a steady state as $t \uparrow + \infty$ with a finite mean value.

(4) The function $z_0(x,t)$ given in (3.7) is a particular delta function initial condition solution of the generalized Feller equation (3.1), (3.3), (3.4). The delta function initial condition solution of this equation with the delta function applied at $x = y > 0$ and $t = 0$ is given by

$$v^*(x,t;y) = (1-\lambda)b^{-1}\xi^{-(p+\lambda)/2} (e^{\beta t \eta})^{(p-\lambda)/2} I_q \left(2\xi^{(1-\lambda)/2} (e^{-\beta t \eta})^{(1-\lambda)/2} \right) \\ \times \exp \left(-\xi^{1-\lambda} - (e^{-\beta t \eta})^{1-\lambda} \right), \quad (3.8)$$

$\xi = xb^{-1}$, $\eta = yb^{-1}$, $b = b(t)$ given by (3.6), $x > 0$, $t > 0$, $q = (\lambda-p)(1-\lambda)^{-1}$, I_q = modified Bessel function of the first kind (Bessel function of imaginary argument). This fact has been established in [9]. (It is useful in this context to also consult [11] and [12] for slight notational differences between this paper and [9].

The function $v^*(x,t;y)$ has the following properties [9]:

- (a) $v^*(x,t;y) > 0$, $x > 0$, $t > 0$, $y > 0$,
- (b) $v^*(x,t;y) \downarrow 0$ as $t \downarrow 0$ for $x > 0$, $y > 0$, $x \neq y$,
- (c) $v^*(x,t;x) \uparrow +\infty$ as $t \downarrow 0$, $x > 0$,
- (d) $v^*(x,t;y) \rightarrow z_0(x,t)$ as $y \downarrow 0$ for $x > 0$, $t > 0$,
- (e) $\int_0^\infty v^*(x,t;y) dx \equiv 1$.

Clearly, these properties make the function $v^*(x,t_0;y)$ a one-sided probability density function for $t=t_0 > 0$ and $y > 0$ fixed. In particular, property (d) substantiates the claim made in the summary that the family of distribution

characterized by (1.3) is a special case of the much more general family specified by $v^*(x; t_0; y)$.

In statistical distribution fitting attempts, in particular in cases where the density data have a maximum, one reason for the frequent occurrence of unsatisfactory fits results from the fact that the location of the maximum of a distribution candidate cannot be chosen arbitrarily. It is normally automatically determined by the basic parameters. For the density functions given by (1.3), for example, the maximum is located at

$$x_m = [-p/(1-\lambda)]^{1/(1-\lambda)} b, \quad p < 0.$$

It is fixed once the parameters b , p , and λ have been determined. The class of functions $v^*(x, t_0; y)$ contains the additional independent "delta function application parameter" y . The presence of this additional parameter changes the situation drastically and favorably. A thorough discussion of the class $v^*(x, t_0; y)$, however, will not be attempted here. We return to the discussion of our main subject.

4. An Applicability Criterion

Inherent in any attempt to fit given empirical distribution data by means of an analytically defined probability density function are three crucial problems, namely (i) candidate function selection from a group of available functions, (ii) determination (estimation) of the parameters of the selected function, and (iii) evaluation of the achieved quality of fit. Since an adequate treatment of the last two problems requires a thorough discussion of the details of the numerical techniques involved they shall be left untouched here. This subject - relative to the class of distributions which represent the topic of the present paper - will be picked up in a separate publication. We shall concentrate, therefore, on the first problem and present a general applicability criterion for the distribution class defined by the cumulative distribution functions (1.1) or by the associated density functions (1.3). This criterion covers all special cases mentioned in Section 1.

Let us consider the distribution function $F(x)$ given in the form (1.2), i.e.,

$$F(x) = \frac{1}{\Gamma(2+q)} \xi^{1-p} \Phi(1+q, 2+q; -\xi^{1-\lambda}), \quad \xi = xb^{-1}.$$

Taking logarithms, we obtain

$$\log F(x) = -\log \Gamma(2+q) + (1-p) \log \frac{x}{b} + \log \Phi\left(1+q, 2+q; -\left(\frac{x}{b}\right)^{1-\lambda}\right). \quad (4.1)$$

At this point it will be advantageous to perform the independent variable transformation $x = \mu y$ where $\mu = m_1$ is the mean (first moment) which can easily be determined from given empirical data. This transformation ensures that all x data in the interval $0 < x < \mu$ will be mapped into y data in the interval $0 < y < 1$. This is important as will become apparent momentarily. Setting then $\log F(x) = \log F(\mu y) = v$ and $\log y = u$ so that

$$\log \frac{x}{b} = \log \frac{y}{b/\mu} = u - \log \mu^{-1}b ,$$

we obtain from (4.1) the functional relation

$$v(u) = -\log \Gamma(2+q) - (1-p)\log \mu^{-1}b + (1-p)u + \log \Phi \left(1+q, 2+q; - \left(\frac{e^u}{\mu^{-1}b} \right)^{1-\lambda} \right). \quad (4.2)$$

The degenerate hypergeometric function Φ is defined as a power series in its last argument with constant term equal to unity. Therefore, as $x \downarrow 0$, i.e., as $y \downarrow 0$ which means as $u \downarrow -\infty$,

$$\log \Phi \left(1+q, 2+q; - (e^u/\mu^{-1}b)^{1-\lambda} \right) \uparrow 0.$$

Consequently, the function $v(u)$ given in (4.2) is asymptotically linear in u as $u \downarrow -\infty$. In other words,

$$v(u) \sim v_1(u) = (1-p)u - \log \Gamma(2+q) - (1-p) \log \mu^{-1}b, \quad u \downarrow -\infty.$$

This asymptotic linearity property may also be expressed by saying that, as $u \downarrow -\infty$, the graph of the function $v(u)$ defined in (4.2) approaches the (straight line) asymptote defined by the linear equation

$$v_1(u) = (1-p)u - \log \Gamma(2+q) - (1-p) \log \mu^{-1}b. \quad (4.3)$$

Based on this fact we can formulate the following Applicability Criterion.
A distribution function $F(x)$ of the class (1.1) may be considered as a candidate for a data fit if the logarithmic plot of a given set of empirical cumulative distribution data indicates the existence of an asymptote.

Remarks. (1) An applicability criterion similar to the one expressed above for the logarithm of the cumulative distribution data can, of course, be formulated for the corresponding density data according to (1.3). Which of these two equivalent criteria is actually being used is immaterial. The one given in terms of the cumulative data is generally preferred simply because the cumulative data are normally "smoother" than the corresponding density data.

(2) An asymptotic linearity criterion similar to the one expressed above for the distribution class (1.1) holds for the class of distributions defined by the density function $v^*(x, t_0; y)$ given in (3.8). This is easily seen. If we denote the cumulative distribution function associated with $v^*(x, t_0; y)$ by $V(x)$, then

$$V(x) \sim F(x) \exp - (e^{-\beta t_0})^{1-\lambda} \quad \text{as } x \downarrow 0$$

where $F(x)$ is given by (1.1). We shall not go into any details here.

There is important practical utility associated with the applicability criterion. This becomes evident when we realize that it can be used to determine approximate values p_1 and λ_1 for the two shape parameters p and λ . An approximate value b_1 for the scale parameter b can then be determined by means of the first moment,

$$b = \mu \frac{\Gamma(1+q)}{\Gamma(1+q+1/(1-\lambda))} \quad (4.4)$$

if we substitute in $q = (\lambda - p)(1 - \lambda)^{-1}$ the values p_1 and λ_1 for p and λ , respectively.

If a set of empirical distribution data indicates the existence of an asymptote for the logarithmic cdf graph, the location of the asymptote can be estimated either by visual inspection or by analytic methods. Numerical techniques for the asymptote determination and for the subsequent estimation of parameters will be discussed elsewhere. The location of the asymptote can be specified by its directional angle ϑ and its intersection with the v -axis. Since the asymptote is determined by the linear equation (4.3), we immediately see that

$$\tan \vartheta = 1 - p. \quad (4.5)$$

This relation makes it possible to quickly find an approximate value p_1 for the initial shape parameter p once ϑ or $\tan \vartheta$ have been estimated,

$$p_1 = 1 - \tan \vartheta.$$

It is of interest to note that, according to (4.5), the principal value of ϑ is uniquely determined by the initial shape parameter p and vice versa. Since $p < 1$, we have $0 < \vartheta < \pi/2$. Some of the distributions listed as special cases in Section 1 have very specific $\tan \vartheta$ values. For the Gauss and exponential distributions we have $p = 0$ so that $\tan \vartheta = 1$. For the Rayleigh distribution $p = -1$ which means that $\tan \vartheta = 2$. For the Maxwell case $p = -2$, $\tan \vartheta = 3$, and in the case of the Wien distribution $p = -3$ so that $\tan \vartheta = 4$.

Next, once the v -axis intercept $v_1(0)$ of the asymptote has approximately been determined, we obtain from (4.3) the equation

$$-\log \Gamma(2+q) - (1-p) \log \mu^{-1} b - v_1(0) = 0.$$

We eliminate the unknown scale parameter b by means of (4.4) which leads to the equality

$$-\log \Gamma(2+q) - (1-p) \log \Gamma(1+q) + (1-p) \log \Gamma(1+q+(1-\lambda)^{-1}) - v_1(o) = 0. \quad (4.6)$$

The left-hand side becomes a function of the unknown terminal shape parameter λ if we replace p by the previously determined approximate value p_1 . In other words, we obtain from (4.6) an equation of the form $\phi(1-\lambda) = 0$. It can be shown that it has exactly one solution $1-\lambda_1 > 0$ (provided $v_1(o)$ has been properly determined) which can easily be obtained by means of Newton's method.

The opportunity to determine "good" approximate values p_1 and λ_1 for the shape parameters p and λ is extremely important for the practical application of the distribution class (1.1). The approximate values p_1 and λ_1 can be used as initial values in any of the established parameter estimation techniques such as, for example, the method of moments or the maximum-likelihood method. Each of these methods leads to three equations for the unknown parameters b , p , and λ . Actually, only two equations are needed since the scale parameter b can be eliminated. The use of the initial values p_1 and λ_1 results in rapid convergence of the iteration process which will lead to the desired final parameter values.

Although the class of probability distributions discussed in this paper has been known for more than sixty years, its application has been limited, most likely as a consequence of computational intensity and possible convergence problems. In general, however, it is not really the complexity of the system of transcendental equations which makes the numerical problem computationally intensive but rather a poor choice initial iteration values. It is hoped that the approach presented here will lead to more widespread use of the distribution class (1.1).

5. Empirical Examples

In the talk presented at the Madison conference two examples based on empirical data have been discussed. As indicated at the beginning of Sec. 4 a thorough treatment of practical examples will not be attempted in this paper. Suffice it, therefore, to simply present the illustrative documentation for the two parameter estimates.

The empirical data were available in histogram (pdf) form as shown in the first figure of each of the two sets of illustrations. The cdf data were obtained by numerical integration. Their logarithmic plots are shown in the second figures, $x_M = \mu$ being the mean. The asymptote data $\tan \theta$ and $v_1(0)$ were determined by visual inspection to obtain approximate values p_1 and $\beta_1 = 1 - \lambda_1$ for the two shape parameters. To improve the numerical values of these parameters the method of moments was used which led to the final values given in the table. The scale parameter b is determined by $b = \mu\theta$, $\mu = \text{mean}$. The last pair of figures show the histograms overlaid with the fitted probability density functions.

CASE #1: HUNTSVILLE AL, DAILY TEMP.RANGE, ANNUAL DISTRIBUTION

Empirical Distribution

No.of Obs.: 9641
 Mean: 21.378 (1)
 Stdd.Dev.: 7.381 (1.1192)
 3rd. Mom.: 16.562 (1.3593)

Fitted Distrib. (Moments' Fit)

P= -1.974
 Beta= 3.584
 Theta= 1.195

CASE #2: HUNTSVILLE AL, RAINFALL INTENSITY, WINTER DISTRIB.

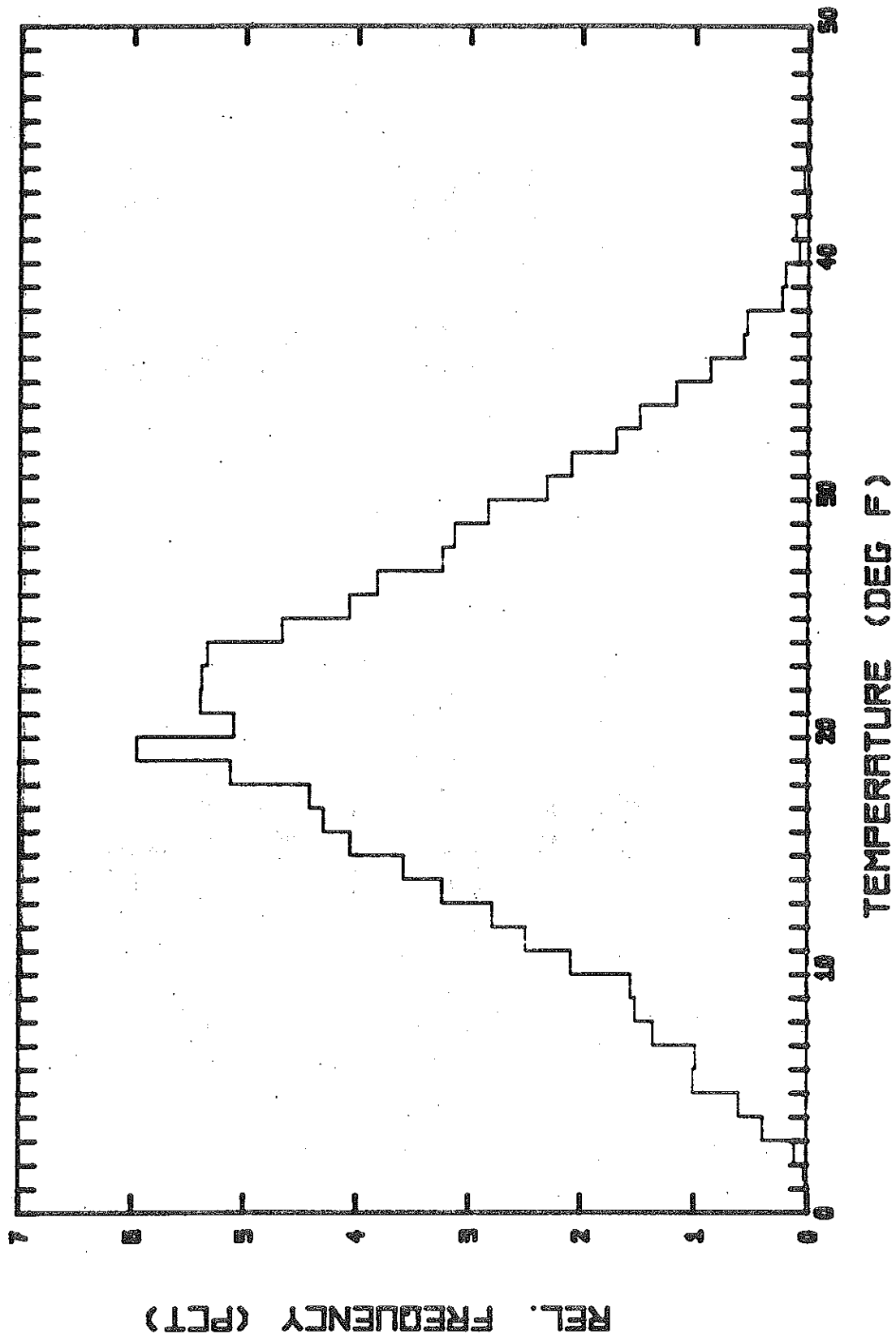
Empirical Distribution

No.of Obs.: 9998
 Mean: 11.736 (1)
 Stdd.Dev.: 6.392 (1.2966)
 3rd. Mom.: 131.705 (1.9714)

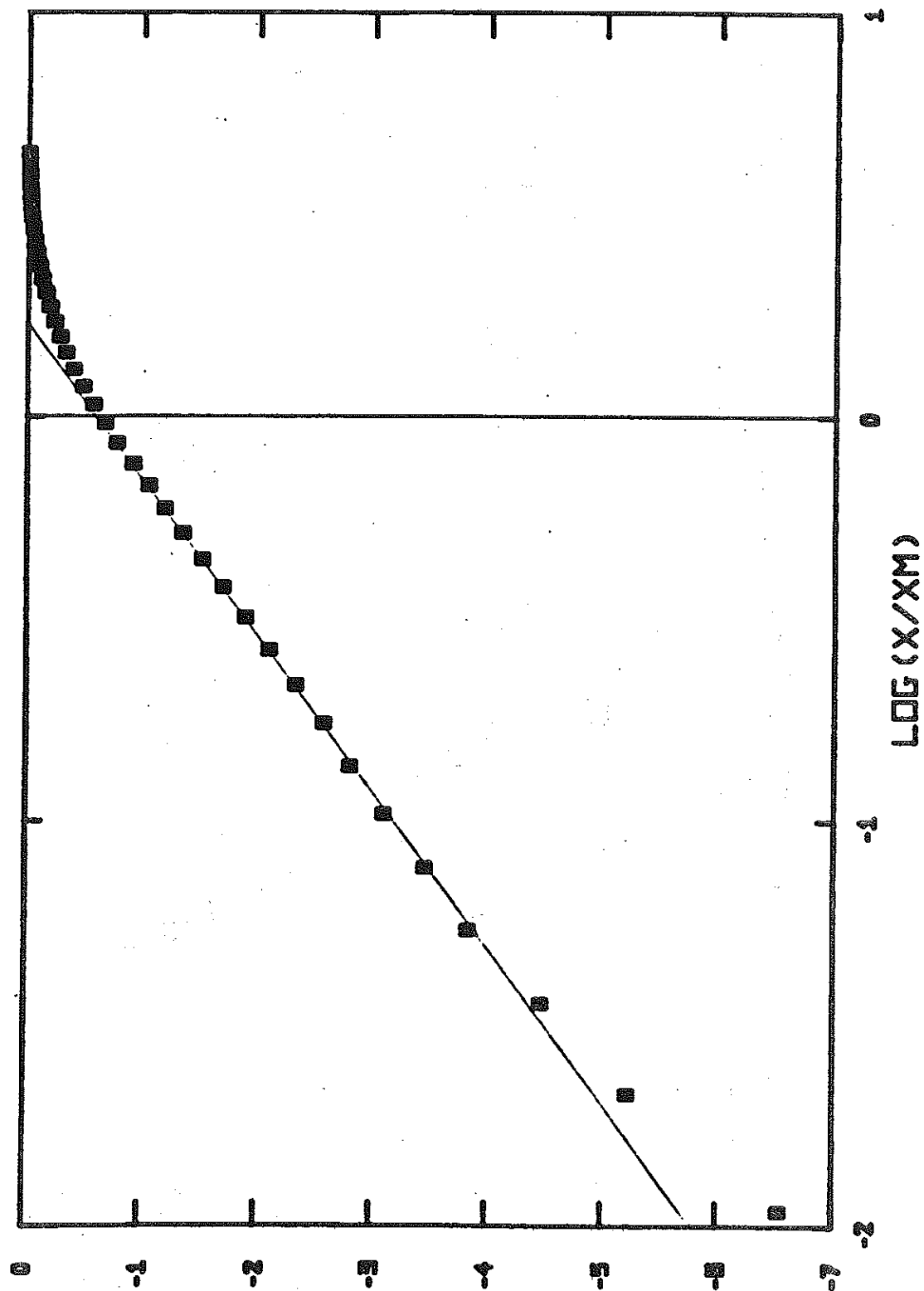
Fitted Distrib. (Moments' Fit)

P= -0.584
 Beta= 2.585
 Theta= 1.460

CASE #1: HSV DAILY TEMP. RANGE ANNUAL DISTRIB.

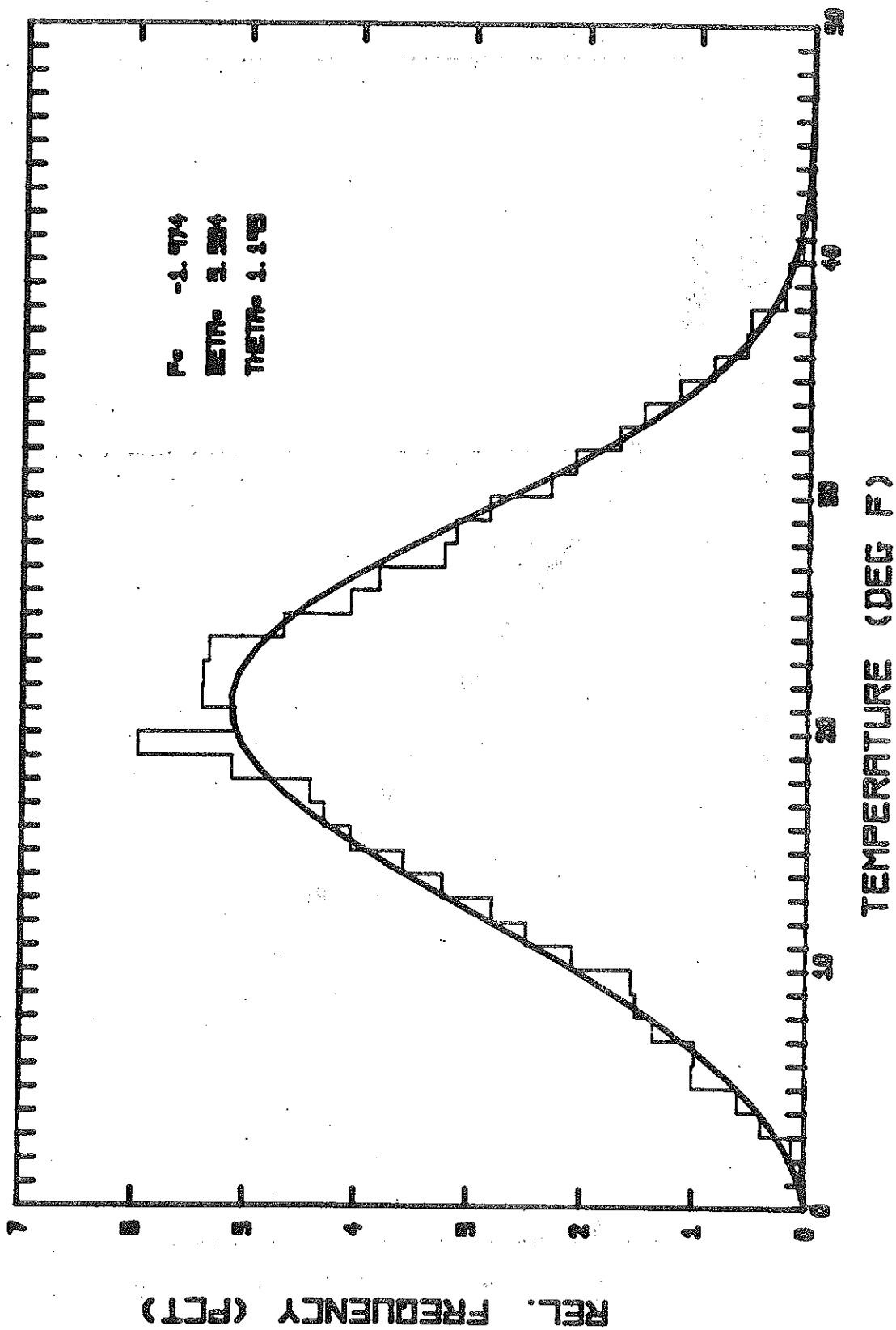


CASE #1: HSV. DAILY TEMP. RANGE. ANNUAL DISTRIBUTION

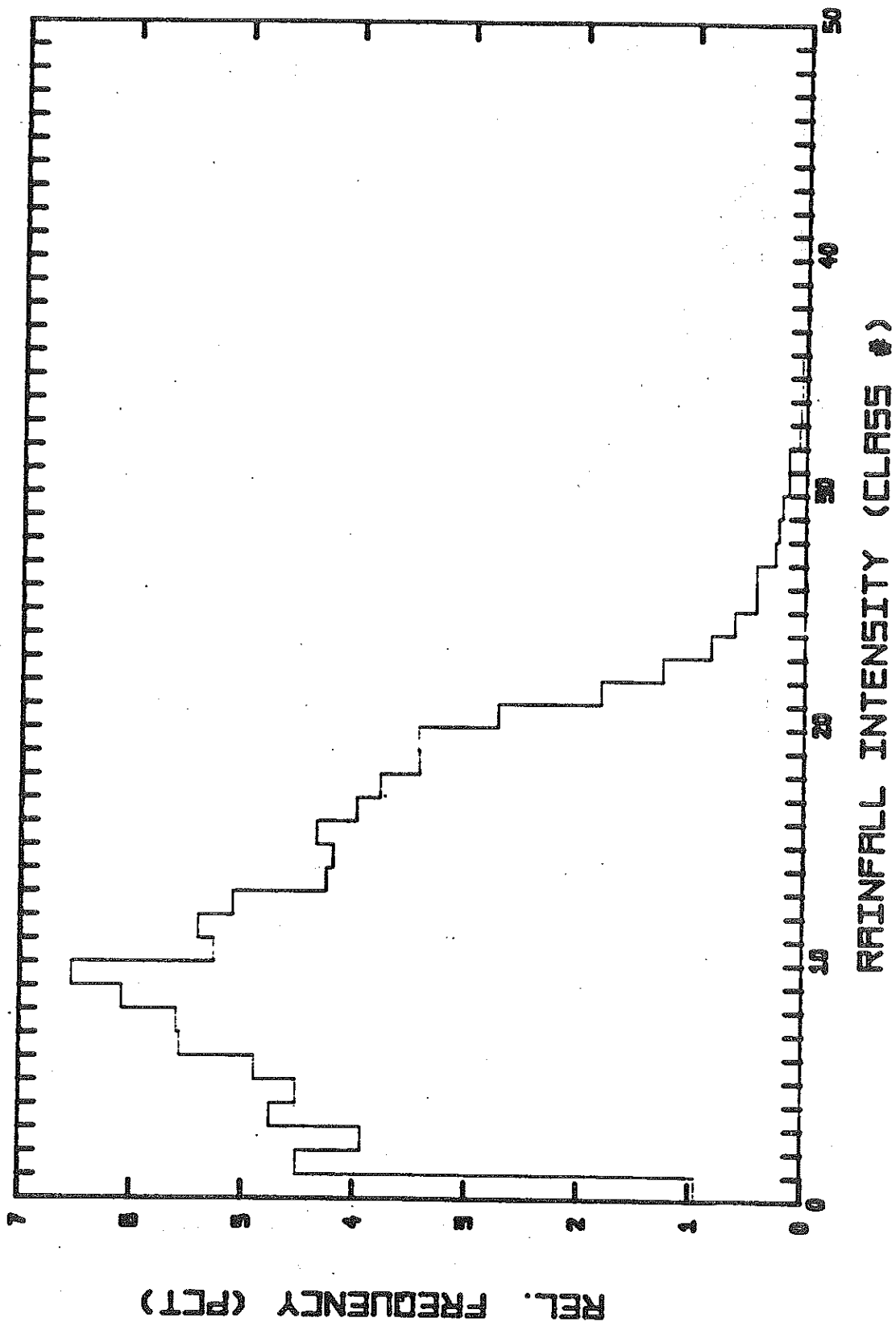


רספ (בחדאחר)

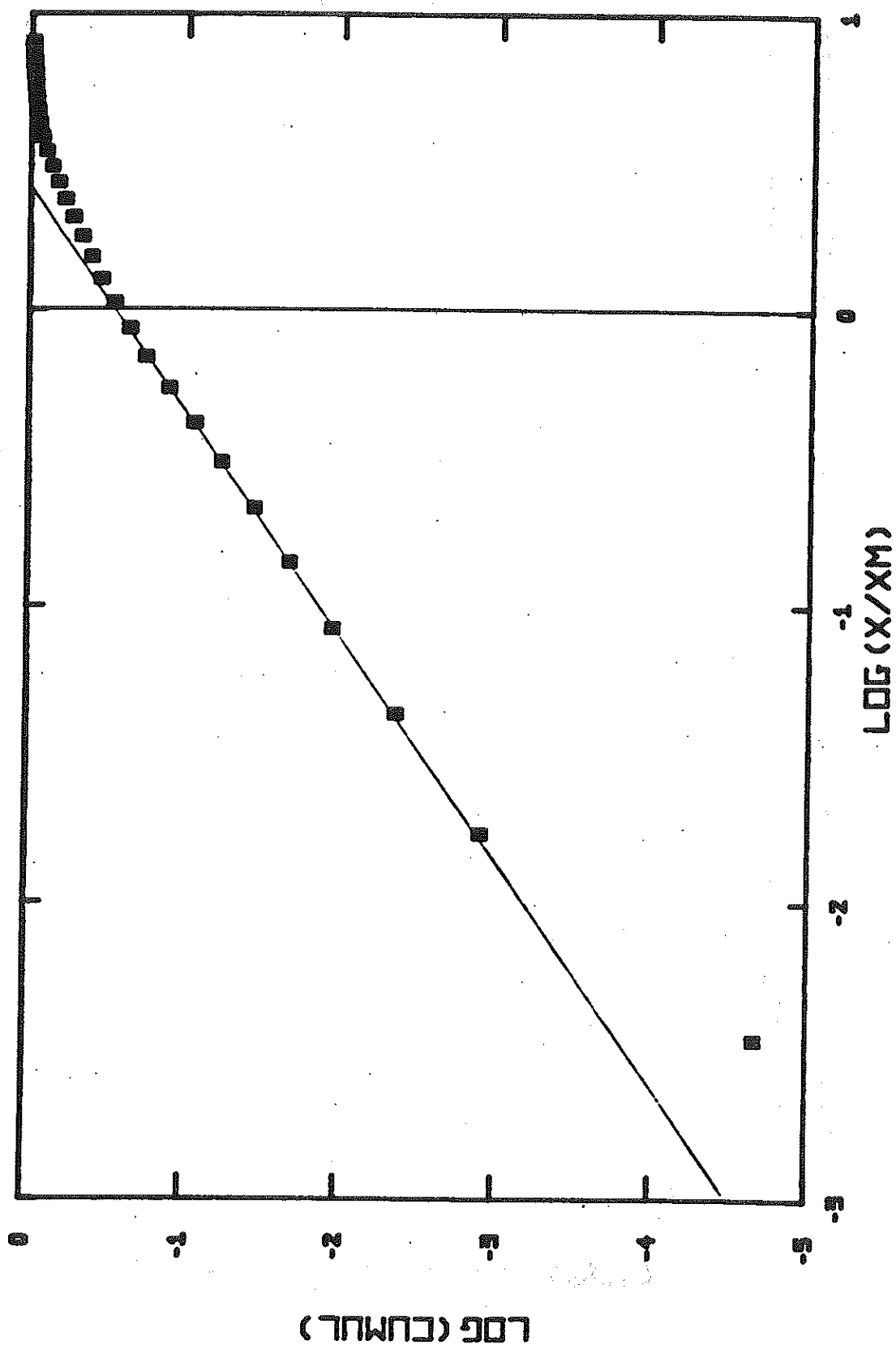
CASE #1: HSV DAILY TEMP. RANGE ANNUAL DISTRIB.



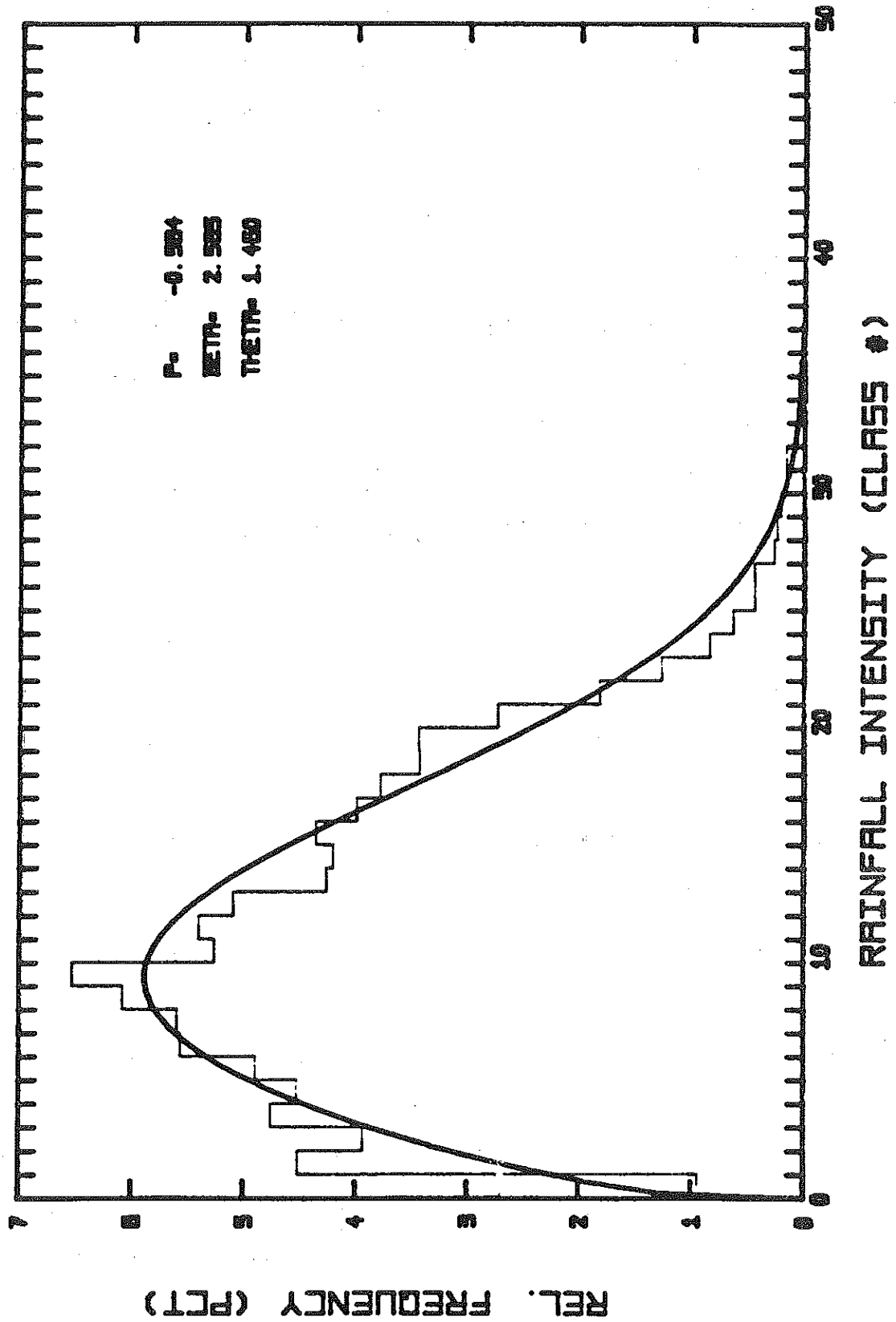
CASE #2: HSV RAINFALL INTENSITY WINTER DISTRIB.



CASE #2: HSV RAINFALL INTENSITY WINTER DISTRIB.



CASE #2: HSV RAINFALL INTENSITY WINTER DISTRIB.



References

- [1] I. S. Gradshteyn, and I. M. Ryzhik, Tables of Integrals, Series, and Products, 4th ed., Academic Press, New York, 1965.
- [2] L. Amoroso, Ricerche intorno alla curva dei redditi, Ann. Mat. Pura Appl., Ser. 4, 2 (1924) 123-157.
- [3] E. W. Stacy, A generalization of the Gamma distribution, Ann. Math. Stat. 33 (1962) 1187-1192.
- [4] Van B. Parr, and J. T. Webster, A method for discriminating between failure density functions used in reliability predictions, Technometrics 7 (1965) 1-10.
- [5] H. L. Harter, Maximum-likelihood estimation of the parameters of a four-parameter generalized Gamma population from complete and censored data, Technometrics 9 (1967) 159-165.
- [6] O. M. Essenwanger, Applied Statistics in Atmospheric Science, Elsevier, Amsterdam, 1976.
- [7] S. H. Lehnigk, A class of probability density functions, submitted, Math. Meth. in the Appl. Sci.
- [8] S. H. Lehnigk, Characteristic function of a class of probability distribution, in preparation.
- [9] S. H. Lehnigk, Initial condition solutions of the generalized Feller equation, J. Appl. Math. Phys. (ZAMP) 29 (1978) 273-294.
- [10] S. H. Lehnigk, Boundary condition solutions of the generalized Feller equation, J. Math. Phys. 19 (1978) 1267-1275.
- [11] S. H. Lehnigk, Biorthogonal sequences of solutions of the generalized Feller equation, Math. Meth. in the Appl. Sci. 4 (1982) 317-353.
- [12] S. H. Lehnigk, A generalized Jacobi Theta function, Math. Meth. in the Appl. Sci. 6 (1984) 327-344.

PLOTTING MATHEMATICAL FUNCTIONS ON A STANDARD LINE PRINTER

DONALD W. RANKIN
Lieutenant Colonel
US Air Force Retired

INTRODUCTION. Often the analyst will be greatly aided if he can view a graph of the function or data under investigation. The wide availability of computer-driven printers suggests that they be adapted to this usage. However, since that is not their primary purpose, some programming is required to exact an acceptable performance from them. This paper, then, discusses some of the principles which must be adhered to and offers some example programs.

No attempt can be made to cover all possible printer-computer combinations, since their number approaches the astronomical. (A recent issue of a periodical lists 145 low- and medium-priced printers from 36 different manufacturers which are compatible with the author's computer!) Instead, a typical combination*, is put forward as an example.

Programming language will be confined to the most elementary BASIC, so that even the casual programmer will feel comfortable. The commands CALL, PEEK, and POKE will not be used. There is little need for streamlining, since even a clumsy program will run faster than the printer.

TYPES OF PRINTERS. The principles herein can be applied to virtually all printers, whether dot matrix, daisy wheel, ink jet or thermal ribbon. Another criterion will be used to roughly divide printers into three categories.

The first type possesses a resident plotting function. For them, this paper is not necessary, although it may contribute some insight.

*An Epson model FX-80 printer driven by a Radio Shack model 100 portable computer.

The second type is capable of a variable reverse line feed. The principal example program is written for this type.

The third type has neither of the above attributes. As will be seen, plotting still may be possible.

SENDING INFORMATION TO THE LINE PRINTER. Most computers send intelligence to the printer as a stream of 8-bit binary numbers (00000000 to 11111111). This corresponds to 0-255 (decimal) or 00-FF (hexadecimal). Some computers send only 7 bits of data, reserving the eighth bit for a parity check or other special use. They cannot distinguish 0xxxxxxx from 1xxxxxxx. This amounts to subtracting 128 wherever possible.

THE CHARACTER-STRING FUNCTION. One means by which BASIC converts information into suitable form is the character-string function, which is implemented by CHR\$(n), where n can vary from 0 to 255. Values of n from 32 to 127 are used to send various symbols, including punctuation, numbers, and all the letters of the alphabet. For example, CHR\$(65) sends a capital A. Values from 0 to 31 are used to send instructions to the various peripherals, and are called control codes. CHR\$(27) is called the ESCAPE code. It alerts the peripheral that one or more binary numbers are to follow, and that the sequence is to be treated as an entity. By using ESCAPE sequences, the number of possible control codes becomes almost unlimited.

Another method of converting to binary is to enclose the actual symbols within quotation marks. Thus LPRINT "A" and LPRINT CHR\$(65) are equivalent. This latter method depends upon the existence of the appropriate symbol, and hence cannot be used to transmit control codes. Also it cannot be used to send actual quotation marks, since BASIC only recognizes them as a sort of switch which turns a binary converter on and off. CHR\$(34) must be used.

Many software designers "borrow" one or more little-used control codes, diverting them

to special uses. When, in running a program, one of them occurs by chance, the result can be most unexpected (and quite unwanted). It is necessary to identify these anomalies, so that the program can avoid them.

THE HEX DUMP. The easiest way to examine the information which the computer is transmitting to the printer is to perform a HEX dump. The printer is placed in hexadecimal mode and the following program executed:

```
10 FOR N = 0 TO 255
20 LPRINT CHR$(N);
30 NEXT N
40 END
```

The resulting printout will identify the codes in question. Note the semicolon at the end of line 20. It inhibits the carriage return. Figure 1 gives an example of a HEX dump.

Figure 1
Radio Shack Model 100 HEX Dump

00	01	02	03	04	05	06	07	08	20	20	20
20	20	20	20	20	0A	0B	0C	0D	0E	0F	10
11	12	13	14	15	16	17	18	19	1B	1C	1D
1E	1F	20	21	22	23	24	25	26	27	28	29
2A	2B	2C	2D	2E	2F	30	31	32	33	34	35
36	37	38	39	3A	3B	3C	3D	3E	3F	40	41
42	43	44	45	46	47	48	49	4A	4B	4C	4D
4E	4F	50	51	52	53	54	55	56	57	58	59
5A	5B	5C	5D	5E	5F	60	61	62	63	64	65
66	67	68	69	6A	6B	6C	6D	6E	6F	70	71
72	73	74	75	76	77	78	79	7A	7B	7C	7D
7E	7F	80	81	82	83	84	85	86	87	88	89
8A	8B	8C	8D	8E	8F	90	91	92	93	94	95
96	97	98	99	9A	9B	9C	9D	9E	9F	A0	A1
A2	A3	A4	A5	A6	A7	A8	A9	AA	AB	AC	AD
AE	AF	B0	B1	B2	B3	B4	B5	B6	B7	B8	B9
BA	BB	BC	BD	BE	BF	C0	C1	C2	C3	C4	C5
C6	C7	C8	C9	CA	CB	CC	CD	CE	CF	D0	D1
D2	D3	D4	D5	D6	D7	D8	D9	DA	DB	DC	DD
DE	DF	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9
EA	EB	EC	ED	EE	EF	F0	F1	F2	F3	F4	F5
F6	F7	F8	F9	FA	FB	FC	FD	FE	FF	0D	

Referring to Figure 1, it can be seen that CHR\$(9) sends a series of spaces, while CHR\$(26) transmits nothing at all. It will be necessary to program around these two values.

SCALING THE PLOT. Some daisy wheel printers have adjustable horizontal spacing. Dot matrix printers achieve somewhat the same effect by offering a selection of type faces. The ability to adjust vertical spacing varies widely. As a rule of thumb, assign the coarser scale factor to the independent variable.

Figure 2

Dot Matrix Type Faces

Type Face	Characters per 6-inch line
Pica	60
Elite	72
Compressed	103

If a printer is capable of reverse line feeds, it is possible to scale and label the plotting area, then return the carriage and platen to a known position before beginning the actual plot. Without this capability, it is necessary to mark the paper in some way so that the platen can be correctly repositioned manually.

PLANNING A PLOTTING PROGRAM. As an exercise, let us write a program which plots two functions simultaneously, using different plotting symbols for each, so that they may be distinguished. Let us assume a dot matrix printer capable of compressed type face and variable reverse line feed. Further let us assume a computer which diverts 09 and 1A (hex) to special uses. (We recall that these codes are generated by CHR\$(9) and CHR\$(26), respectively.) Available plotting area is 6 by 6 inches.

To generate variable line feeds, the ESCAPE sequences 1B;4A;nn (forward) and 1B;6A;nn (reverse) are used, where nn can vary from 00 to FF (except 09 and 1A, of course). BASIC uses CHR\$(27)CHR\$(74)CHR\$(N) and CHR\$(27)CHR\$(106)CHR\$(N) to send these sequences (N = 0 to 255). Since symbols exist for CHR\$(74) and CHR\$(106), the shorter forms CHR\$(27)"J"CHR\$(N) and CHR\$(27)"j"CHR\$(N) can be used. Some computers may require semicolons between the parts. For the printer which was employed, a value of N = 255 moves the platen exactly 3 cm. Thus there are 85 machine counts per cm., or 216 per inch.

By using the compressed type face for plotting, we find 27 machine counts per 4 cm., or 103 per 6 inches. It is apparent at once that the independent variable should vary in the horizontal direction.

For an example plot, choose the tangent and cosine functions through the range from 0 to 240 degrees, inclusive. Assigning a scale factor of 2.5 degrees per character, the plot will be 97 characters wide (compressed), which leaves a few for labelling. The computer requires that the argument be expressed in radians, so that one character is equivalent to 0.0436332313 radians. Successive values of the functions are computed by a routine similar to:

```
10 FOR X = 0 TO 96
20 C = COS(0.0436332313 * X)
30 IF ABS(C) < 0.3 THEN 50
40 T = TAN(0.0436332313 * X)
50 NEXT X
```

Line 30 is not essential. It merely avoids computing large values of the tangent which would not be plotted anyway.

For the vertical scale, let us choose unity to be 1.25 inches. Now the ordinates can be easily read with a common foot ruler, since 0.1 = 1/8". Multiplying 216 by 1.25, it is found that there are 270 machine counts per unit on the vertical axis. Values to about ± 2.15 can be displayed within the allotted area.

The number to be converted to binary by the character-string function must be an integer. This can be accomplished by:

$$Y = \text{INT}(.5 + 270 * C)$$

Computed in this way, there is no need to worry about sign.

There is one more point to consider. The program must be given a memory. It is vital that it be able to "remember" the position of the platen. For this express purpose, the variable LY (last Y) is established.

Below is an example program, followed by explanatory notes.

A SAMPLE PLOTTING PROGRAM.

```

2990 END
3000 LPRINT CHR$(27)"1"CHR$(10);LPRINT CHR$(27)"A"
CHR$(8)
3010 J$ = "      |      |      |      |
      |      |      |      |
      |:LPRINT CHR$(15);J$:LPRINT J$
3020 FOR NZ = 1 TO 22
3030 LPRINT "|";TAB(36);"|";TAB(72);|"
3040 NEXT NZ
3050 LPRINT J$
3060 FOR NZ = 1 TO 22
3070 LPRINT "|";TAB(36);"|";TAB(72);|"
3080 NEXT NZ
3090 LPRINT J$:LPRINT J$
3100 LPRINT CHR$(18);"0";TAB(7);"30";TAB(14);"60";
TAB(21);"90";TAB(28);"120";TAB(35);"150";TAB(42);
"180";TAB(49);"210";TAB(56);"240":LPRINT
3110 LPRINT TAB(26);"Degrees";CHR$(15);CHR$(27)"A"
CHR$(0)
3120 K$ = "-----
-----
----- "
3130 LPRINT CHR$(27)"j"CHR$(108);K$;CHR$(18);"-2";
CHR$(15)
3140 LPRINT CHR$(27)"j"CHR$(135);CHR$(27)"j"CHR$(
135);K$;CHR$(18);"-1";CHR$(15)
3150 LPRINT CHR$(27)"j"CHR$(135);CHR$(27)"j"CHR$(
135);K$;CHR$(18);" 0";CHR$(15)

```

```

3160 LPRINT CHR$(27)"j"CHR$(135);CHR$(27)"j"CHR$(
135);K$;CHR$(18);" 1";CHR$(15)
3170 LPRINT CHR$(27)"j"CHR$(135);CHR$(27)"j"CHR$(
135);K$;CHR$(18);" 2"
3180 LPRINT CHR$(27)"J"CHR$(72);TAB(31);"Figure 3"
3190 LPRINT CHR$(27)"J"CHR$(72);TAB(30);"+y = cos x"
3200 LPRINT CHR$(27)"J"CHR$(54);TAB(30);"*y = tan x"
;CHR$(15)
3210 LPRINT CHR$(27)"J"CHR$(171);CHR$(27)"J"CHR$(171)
3220 Y = 0
3230 FOR X = 0 TO 96
3240 C = 270 * COS(0.0436332313 * X) : T = 999
3250 IF ABS(C) < 99 THEN 3270
3260 T = 270 * TAN(0.0436332313 * X)
3270 IF ABS(T) > 580 THEN 3460
3280 LY = Y : Y = INT(.5 + T)
3290 IF Y > LY THEN 3390
3300 IF Y < LY THEN 3320
3310 LPRINT "*";CHR$(8);: GOTO 3460
3320 IF (LY - Y) < 256 THEN 3340
3330 LY = LY - 255 : LPRINT CHR$(27)"J"CHR$(255);:
GOTO 3320
3340 IF (LY - Y) = 26 THEN 3370
3350 IF (LY - Y) = 9 THEN 3380
3360 LPRINT CHR$(27)"J"CHR$(LY-Y);"*";CHR$(8);: GO
TO 3460
3370 LPRINT CHR$(27)"J"CHR$(13);CHR$(27)"J"CHR$(13)
;"*";CHR$(8);: GOTO 3460
3380 LPRINT CHR$(27)"J"CHR$(4);CHR$(27)"J"CHR$(5);
"*"CHR$(8);: GOTO 3460
3390 IF (Y-LY) < 256 THEN 3410
3400 LY = LY + 255 : LPRINT CHR$(27)"j"CHR$(255);:
GOTO 3390
3410 IF (Y-LY) = 26 THEN 3440
3420 IF (Y-LY) = 9 THEN 3450
3430 LPRINT CHR$(27)"j"CHR$(Y-LY);"*";CHR$(8);: GO
TO 3460
3440 LPRINT CHR$(27)"j"CHR$(13);CHR$(27)"j"CHR$(13)
;"*";CHR$(8);: GOTO 3460
3450 LPRINT CHR$(27)"j"CHR$(4);CHR$(27)"j"CHR$(5);
"*";CHR$(8);
3460 LY = Y : Y = INT(.5 + C)
3470 IF LY > Y THEN 3500
3480 IF LY < Y THEN 3570
3490 LPRINT "+";: GOTO 3640
3500 IF (LY-Y) < 256 THEN 3520
3510 LPRINT CHR$(27)"J"CHR$(255);: LY = LY - 255 :
GOTO 3500

```

```

3520 IF (LY-Y) = 26 THEN 3550
3530 IF (LY-Y) = 9 THEN 3560
3540 LPRINT CHR$(27)"J"CHR$(LY-Y);"+";: GOTO 3640
3550 LPRINT CHR$(27)"J"CHR$(13);CHR$(27)"J"CHR$(13)
;"+";: GOTO 3640
3560 LPRINT CHR$(27)"J"CHR$(4);CHR$(27)"J"CHR$(5);
"+";: GOTO 3640
3570 IF (Y-LY) < 256 THEN 3590
3580 LPRINT CHR$(27)"j"CHR$(255);: LY = LY + 255 :
GOTO 3570
3590 IF (Y-LY) = 26 THEN 3620
3600 IF (Y-LY) = 9 THEN 3630
3610 LPRINT CHR$(27)"j"CHR$(Y-LY);"+";: GOTO 3640
3620 LPRINT CHR$(27)"j"CHR$(13);CHR$(27)"j"CHR$(13)
;"+";: GOTO 3640
3630 LPRINT CHR$(27)"j"CHR$(4);CHR$(27)"j"CHR$(5);
"+";
3640 NEXT X
3650 LY = Y : Y = -720
3660 IF (LY-Y) < 256 THEN 3680
3670 LY = LY - 255 : LPRINT CHR$(27)"J"CHR$(255) :
GOTO 3660
3680 IF (LY-Y) = 26 THEN 3710
3690 IF (LY-Y) = 9 THEN 3720
3700 LPRINT CHR$(27)"J"CHR$(LY-Y) : GOTO 3730
3710 LPRINT CHR$(27)"J"CHR$(13);CHR$(27)"J"CHR$(13)
: GOTO 3730
3720 LPRINT CHR$(27)"J"CHR$(4);CHR$(27)"J"CHR$(5)
3730 LPRINT CHR$(27)"2";CHR$(18)
3740 RETURN

```

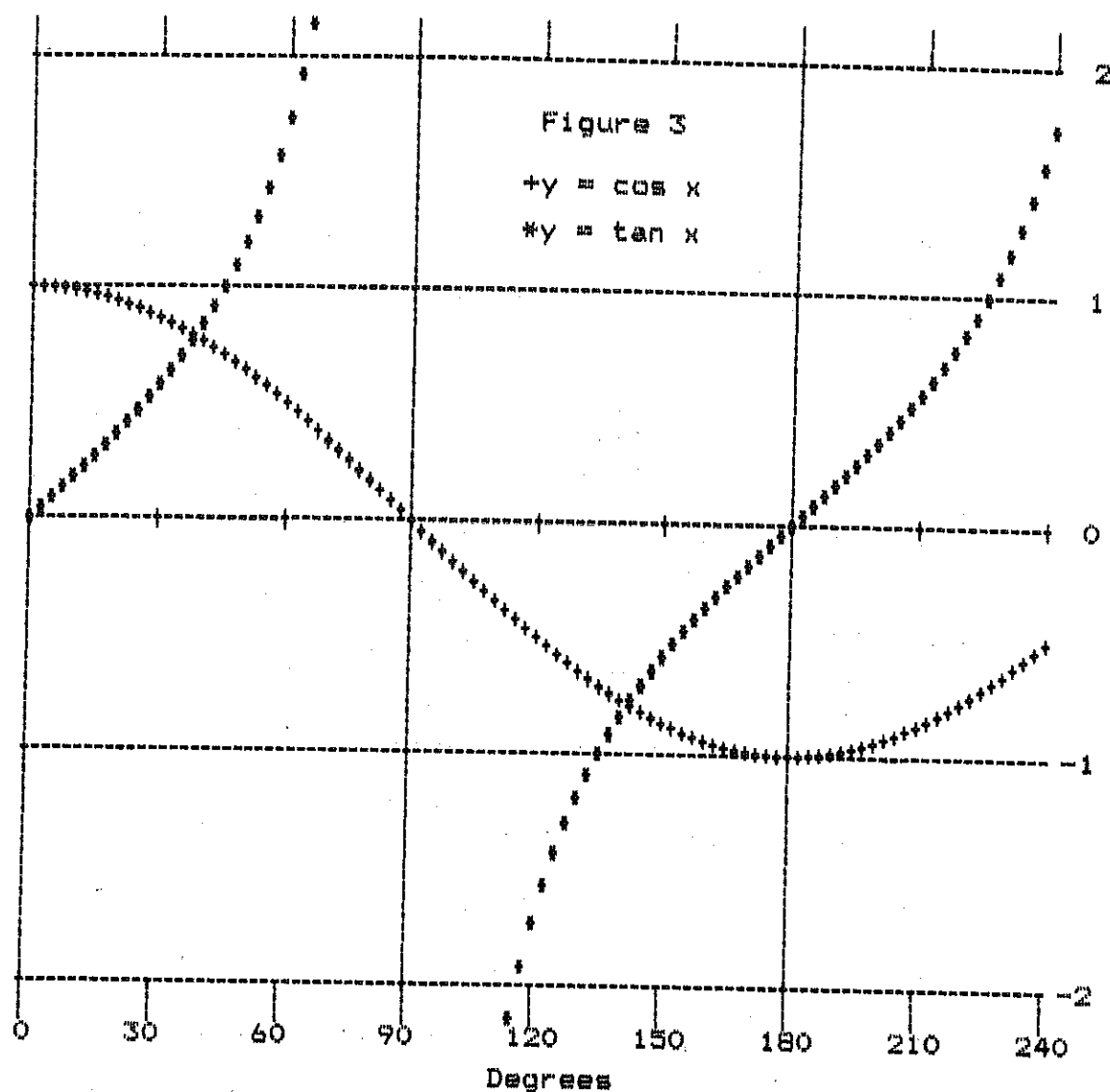
NOTES ON THE PROGRAM.

Line 3000. Sets left margin to 1.25 in. Sets line feed to 1/9 in. for cosmetic purposes. The colon is used to separate statements on the same numbered line. Some computers may require a different symbol.

Lines 3010-3090. Plots the vertical grid. CHR\$(15) calls up the compressed type face. The string variable J\$ must contain a count of 11 spaces between each symbol "!". The symbol is generated by CHR\$(124), or it can be reached from the keyboard with the keystrokes <SHIFT><GRPH><->.

Lines 3100-3200. Plots the horizontal grid, labelling as it goes. Note that the plot is in Compressed type face, but the labelling is done in Pica. The width ratio is 7:12. CHR\$(18) restores Pica.
 Line 3110. CHR\$(15) calls for the Compressed type face. CHR\$(27)"A"CHR\$(0) kills the line feed associated with a carriage return.
 Lines 3210-3220. The platen, carriage, and dependent variable are zeroed.
 Lines 3230-3640. Computation and plotting are accomplished by means of a FOR-NEXT loop.
 Line 3270. This places a limit on the values which will be plotted. Without this limit, the program might attempt to plot a point off the paper, thereby jamming the paper under the platen.
 Line 3310. CHR\$(8) generates a backspace. The trailing semicolon inhibits the carriage return.
 Lines 3320-3330. Moves platen in steps of 3 cm. when required.
 Lines 3340-3380. Moves the platen and plots the point, avoiding the problem codes 9 and 26. This pattern is repeated three times (two functions, two signs).
 Lines 3650-3720. The platen is moved to the bottom of the plot, in position for following text.
 Line 3730. Restores normal line feed and Pica type face.
 Line 3740. If the program is not used as a sub-routine, substitute "END" or "GOTO nnn".

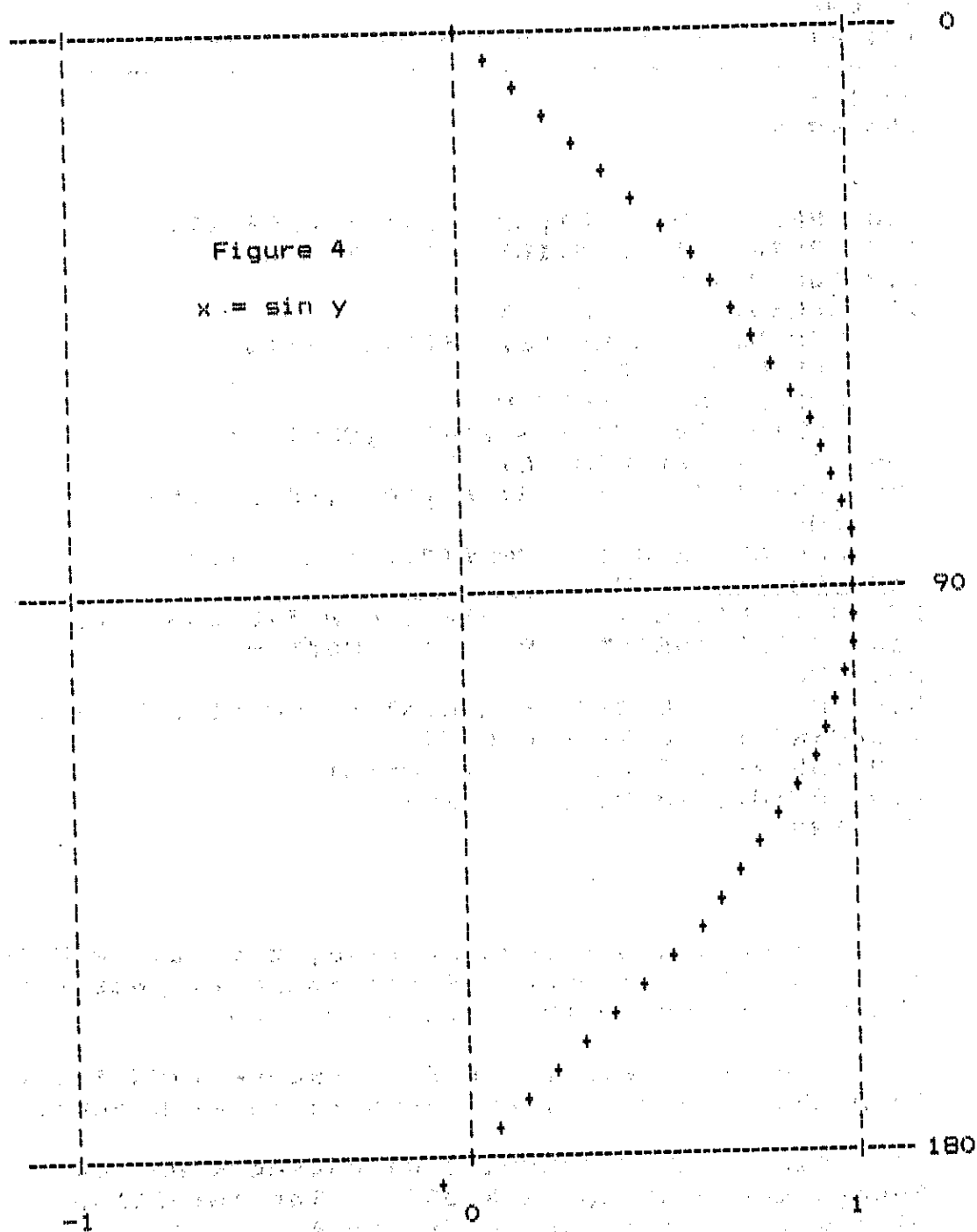
Figure 3 illustrates the program exercised.



PLOTTING WITHOUT VARIABLE REVERSE LINE FEED.

For printers which lack the desired functions, it may suffice to use minimum line feeds to increment the independent variable, and the TAB function to plot the dependent variable. Begin by printing a horizontal line, which is used as a reference mark for aligning the paper with the printer's paper guide bar.

The results of such a technique are shown in Figure 4.



```

5 END
10 LPRINT "-----"
-----"
15 LPRINT : LPRINT : LPRINT : LPRINT : LPRINT
20 END
100 J$ = "-----"
-----"
105 K$ = "      |
      |
      |"
110 LPRINT CHR$(15);CHR$(27)"1"CHR$(17)
115 LPRINT J$;CHR$(18);"0";CHR$(15)
120 FOR N = 1 TO 19
125 LPRINT K$ : NEXT N
130 LPRINT J$;CHR$(18);"90";CHR$(15)
135 FOR N = 1 TO 19
140 LPRINT K$ : NEXT N
145 LPRINT J$;CHR$(18);"180";CHR$(15)
150 LPRINT K$;CHR$(18)
155 LPRINT " -1";TAB(26);"0";TAB(49);"+1"
160 END
200 LPRINT CHR$(15);CHR$(27)"1"CHR$(17)
210 FOR N = 0 TO 41
220 S = INT(.5 + 40 * SIN(0.07853981634 * N))
230 LPRINT TAB(45 + S);"+" : NEXT N
240 END
300 LPRINT : LPRINT : LPRINT : LPRINT : LPRINT
: LPRINT : LPRINT : LPRINT
310 TAB(21);"Figure 4" : LPRINT
320 LPRINT TAB(20);"x = sin y
330 END

```

To draw the reference line, execute <RUN 10>. Then turn the printer off and manually position the platen, using the reference line.

Turn the printer on and execute <RUN 100>. Turn the printer off and reposition as before.

Repeat the procedure executing <RUN 200>. Repeat again using <RUN 300>. The resulting plot will be similar to Figure 4.

STATISTICAL COMPARISON OF THE ABILITY OF
CAMOUFLAGE COLORS TO BLEND WITH TERRAIN BACKGROUND
UNDER HIGH AND LOW SUN ANGLES

George Anitole and Ronald I. Johnson
U.S. Army Belvoir Research and Development Center
Fort Belvoir, Virginia 22060

Christopher J. Neubert
U.S. Army Engineer School
Fort Belvoir, Virginia 22060

ABSTRACT

This study determined the effect of sunlight angle upon the effectiveness of camouflage colors to blend with desert backgrounds. Eleven U.S. Marine personnel and two civilians subjectively evaluated ten colors at nine desert sites, under high and low sunlight angles. The best six colors were rated on a six point scale, with the value number one most effective, and number six not effective. An analysis of variance was performed for each site and all nine sites combined to determine significant ($\alpha = 0.05$) differences between the best four colors. Tukey's Studentized Range Test for Variable Ratings identified which of the four colors differed significantly ($\alpha = 0.05$) from each other. Slight differences were found in the ranking of the colors. This eliminates the requirements for low angle sunlight data.

1.0 SECTION 1 - Introduction

This Center started its current desert color evaluations in April 1980, when the Project Manager, Saudi Arabian National Guard (SANG) Modernization requested camouflage for SANG. Field color evaluations have been conducted in Saudi Arabia and the United States desert southwest. During these studies it was noted that the camouflage colors became brighter in hue when subjected to low sunlight angles in the early morning or late afternoon. This observation led to the question - what effects do high and low sunlight angles have upon the judgment of how well camouflage colors blend with the desert background? This paper presents the results of a study conducted in the United States deserts designed to answer the above question. It should be noted that if testing is required under both high and low sunlight angles, the costs and time to run the study were about doubled. If evaluations can be completed using one sunlight angle, the high sunlight angle would be tested rather than the low sunlight angle, because of its much longer time duration in the course of a day.

2.0 SECTION 2 - Experimental Design

2.1 Camouflage Colors

With the exception of the paint colors Gun Metal Gray and Egyptian, all the colors studied were taken from the SANG color test palette. These

colors were developed over a two-year period, and they represent the most sophisticated available to determine camouflage effectiveness for a series of selected different desert sites. The Gun Metal Gray color was selected to provide high color contrast (in patterns). The Egyptian color is the paint currently being used to camouflage Egyptian equipment. Two new paints derived from the Saudi Arabian desert color palette were colors W and X. Color W is a fifty-fifty mix of colors 7 and 8*, while X is color 11 with the addition of black paint. All paints were lusterless with a reflectance of 1% at a 60° angle.

2.2 Test Targets

The test targets used for this study had to be highly mobile and large enough to permit a study of the target with various desert backgrounds. The U.S. Marine Corps made available ten Commercial Utility Cargo Vehicle (CUCV) trucks which were painted and coded according to Table 1. Each truck was painted on the basis of a three color pattern and are identified as colors 1, 2, and 3. For monotonies and two color patterns, one or more color is repeated.

2.3 Test Sites

A total of nine sites were selected for this study. All the desert sites contained sparse vegetation similar to that found in Saudi Arabia. The soil ranged in color from a light buff/tan to gray and dark brown, and

TABLE 1
CUCV Truck Colors

<u>Vehicle Number</u>	<u>Color</u>		
	<u>1</u>	<u>2</u>	<u>3</u>
A	3	3	3
B	5	3	1
C	7	E*	8
D	7	8	8
F	11	11	11
G	Gun Metal Gray	3	5
H	8	8	8
I	10	10	10
W	7/8	7/8	7/8
X	AC11	AC11	AC11

* Egyptian Color

represented a good cross-sectional spectrum of different colored desert backgrounds. For example, one site on Midland Road, Blythe, California, had a reddish color, while the site at the Baker, California, dry lake was dark brown. The site at Jean Dry Lake bed off Route 15 in Nevada was somewhat yellow in appearance. The order of the nine sites as they will appear throughout this study is seen in Table 2.

*numerical designations were assigned to colors during prior field tests

TABLE 2
Site Order Identification

<u>Site #</u>	<u>Color</u>	<u>Location</u>
1	Buff	Yuma Sand Dunes, AZ
2	Light Gray	Ogilby Road, CA
3	Gray-Tan	Baker Sand Dunes, CA
4	Light Buff/Tan	29 Palms, Range 111, CA
5	Light Tan	29 Palms, Tank Trail, CA
6	Reddish Tan	Midland Road, Blythe, CA
7	Yellow-Tan	Jean Dry Lake Bed, Las Vegas, NV
8	Brown	Dry Lake Bed, Baker, CA
9	Dark Tan	Salton Sea, CA

2.4 Test Subjects

The test subjects consisted of eleven U.S. Marine Corps enlisted men and two civilian employees from the Countersurveillance and Deception Division, Fort Belvoir, Virginia. The enlisted personnel belonged to the 1st Marine Amphibious FORCE Service Support Group, Camp Pendleton, California. Thus, each ground observation consisted of a sample size of thirteen. Each subject had at least a corrected visual acuity of 20/30 and normal color vision.

2.5 Data Generation

The object of this study was to determine what effects high and low sunlight angles have on the ability of camouflage paint colors to blend with desert backgrounds. The relative rating of these colors under the two sunlight conditions was compared to determine significant differences. The ten trucks were painted as shown in Table 1. The trucks were divided into the following two groups:

A B C F W

G H I D X

By using this division, two of the patterned trucks appeared in each of the two groups along with three monotonies. The ground observers (13) were asked to select three color combinations from each of the two groups, based upon their subjective judgment in the colors ability to blend the CUCV trucks with the desert background.

The next task was to rank the remaining six colors on their ability to blend with the desert background using the following ranking system:

- 1 - Most effective
- 2 - Very effective
- 3 - Effective
- 4 - Somewhat effective
- 5 - Less effective
- 6 - Not effective

No ties were allowed. Each of the six colored trucks was assigned a number. A value of 7 was assigned for all colors not selected for final ranking by the ground observers.

3.0 SECTION 3 - Results

The results of each site for both the high and low sunlight angles will not be included because it would be too voluminous to present in these proceedings. A summary of the four best colors for each site under high and low sunlight angles is included in the discussion section. This data is available upon request from the U.S. Army Belvoir Research and Development Center, ATTN: STRBE-JDS, Fort Belvoir, VA 22060. Tables 3-5 and Figure 1 show the data and data analysis averaged across all nine sites for the high sunlight angle. Tables 6-8 and Figure 2 show the data and data analysis averaged across all nine sites for the low sunlight angle. Table 9-11 and Figure 3 show the data and data analysis for the combined high and low sunlight angles to determine what effects high and low sunlight angles had upon the camouflage colors in their ability to blend with the desert background.

TABLE 3

Descriptive Data for CUCV Truck Color Blend with Desert Background, Averaged Across All Sites, High Sunlight Angle

<u>COLOR</u>	<u>N</u>	<u>MEAN</u>	<u>STD ERROR OF COL MEAN</u>	<u>95% CONFIDENCE INTERVAL</u>	
				<u>LOWER LIMIT</u>	<u>UPPER LIMIT</u>
A	117	5.76923	0.218300	5.34136	6.19710
B	117	6.27350	0.150461	5.97860	6.56841
C	117	4.76923	0.158920	4.45775	5.08071
D	117	3.83761	0.124956	3.59269	4.08252
F	117	4.28205	0.146099	3.99570	4.56840
G	117	7.00000	0.000000	7.00000	7.00000
H	117	3.82051	0.142922	3.54039	4.10064
I	117	6.70940	0.088425	6.53609	6.88272
W	117	3.60684	0.217140	3.18124	4.03243
X	117	2.92308	0.190843	2.54902	3.29713

TABLE 4

Analysis of Variance for the Best Four Color Blends, Averaged Across All Sites, High Sunlight Angle

<u>SOURCE</u>	<u>DF</u>	<u>SUM OF SQUARES</u>	<u>MEAN SQUARE</u>	<u>F VALUE</u>	<u>PR>F</u>
COLOR	3	64.59829060	21.53276353	6.15	0.0005
ERROR	464	1623.36752137	3.49863690		
TOTAL	467	1687.96581197			

Table 4 indicates that there are significant differences in the ability of the top four colors to blend with the desert background. These differences are shown in Table 5.

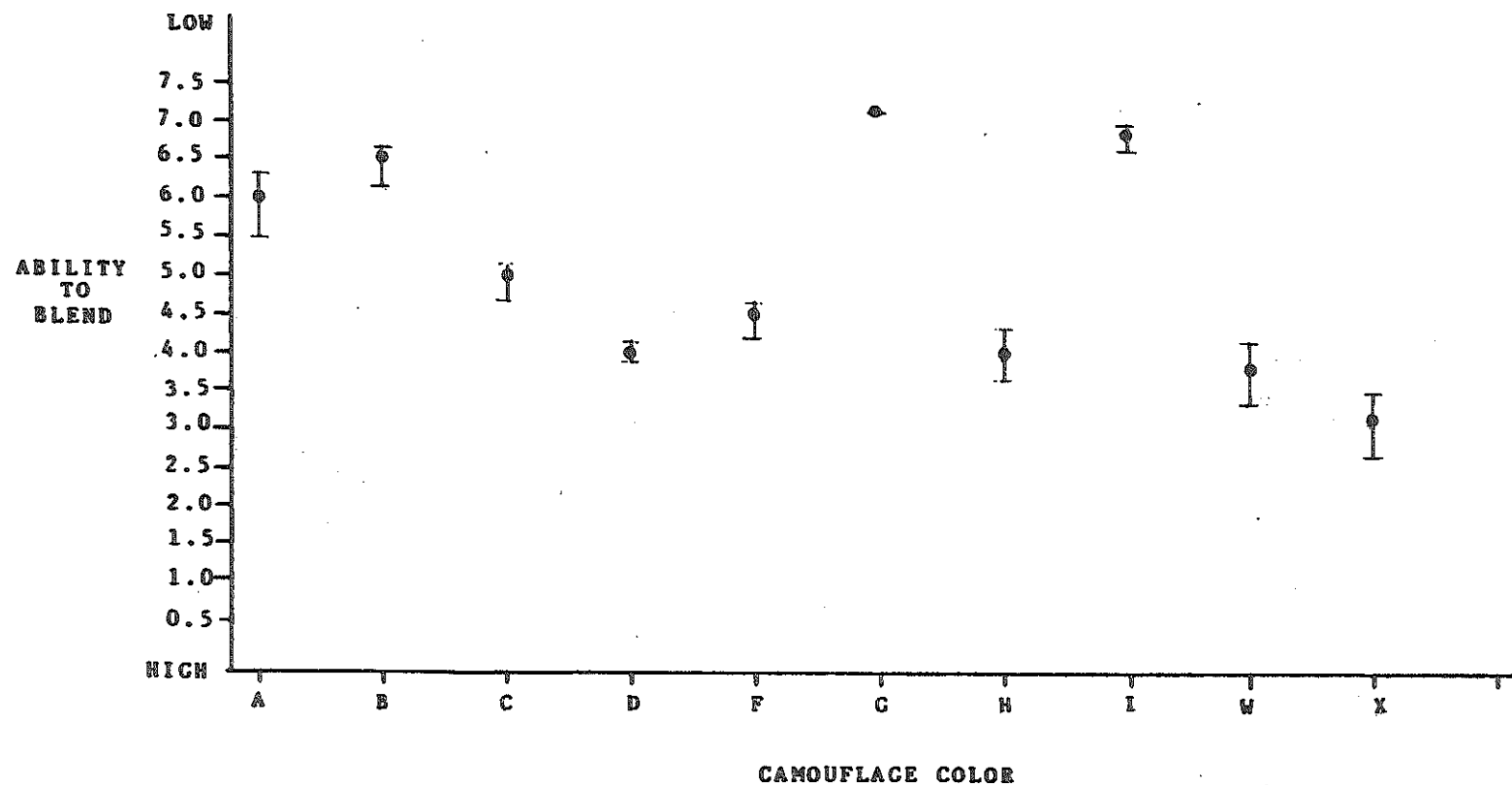


Figure 1 - Camouflage Color Ability to Blend CUCV Truck with Desert Background, Averaged Across All Nine Sites, High Sunlight Angle

TABLE 5

Significant Differences Between the Top Four Camouflage Colors
(Blend), Averaged Across All Sites, High Sunlight Angle

<u>TUKEY GROUPING</u>	<u>MEAN</u>	<u>N</u>	<u>COLORS</u>
A	3.8376	117	D
A	3.8205	117	H
A	3.6068	117	W
B	2.9231	117	X

$\alpha = 0.05$, Degrees of Freedom = 464

Critical Value of Studentized Range = 3.646

Minimum Significant Difference = 0.630546

Color means with the same letter in the grouping column are not significantly different.

TABLE 6

Descriptive Data for CUCV Truck Color Blend with Desert
Background, Averaged Across All Sites, Low Sunlight Angle

<u>COLOR</u>	<u>N</u>	<u>MEAN</u>	<u>STD ERROR OF COL MEAN</u>	<u>95% CONFIDENCE INTERVAL</u>	
				<u>LOWER LIMIT</u>	<u>UPPER LIMIT</u>
A	117	5.76923	0.191703	5.39349	6.14497
B	117	7.00000	0.000000	7.00000	7.00000
C	117	5.31624	0.141385	5.03913	5.59335
D	117	3.81197	0.107850	3.60058	4.02335
F	117	4.21368	0.152988	3.91382	4.51353
G	117	7.00000	0.000000	7.00000	7.00000
H	117	4.18803	0.139961	3.91371	4.46236
I	117	7.00000	0.000000	7.00000	7.00000
W	117	2.19658	0.144268	1.91382	2.47935
X	117	2.50427	0.137675	2.23443	2.77412

TABLE 7

Analysis of Variance for the Best Four Color Blends,
Averaged Across All Sites, Low Sunlight Angle

<u>SOURCE</u>	<u>DF</u>	<u>SUM OF SQUARES</u>	<u>MEAN SQUARE</u>	<u>F VALUE</u>	<u>PR>F</u>
COLOR	3	332.17948718	110.72649573	53.33	0.0001
ERROR	464	963.45299145	2.07640731		
TOTAL	467	1295.63247863			

Table 7 indicates that there are significant differences in the ability of the top four colors to blend with the desert background. These differences are shown in Table 8.

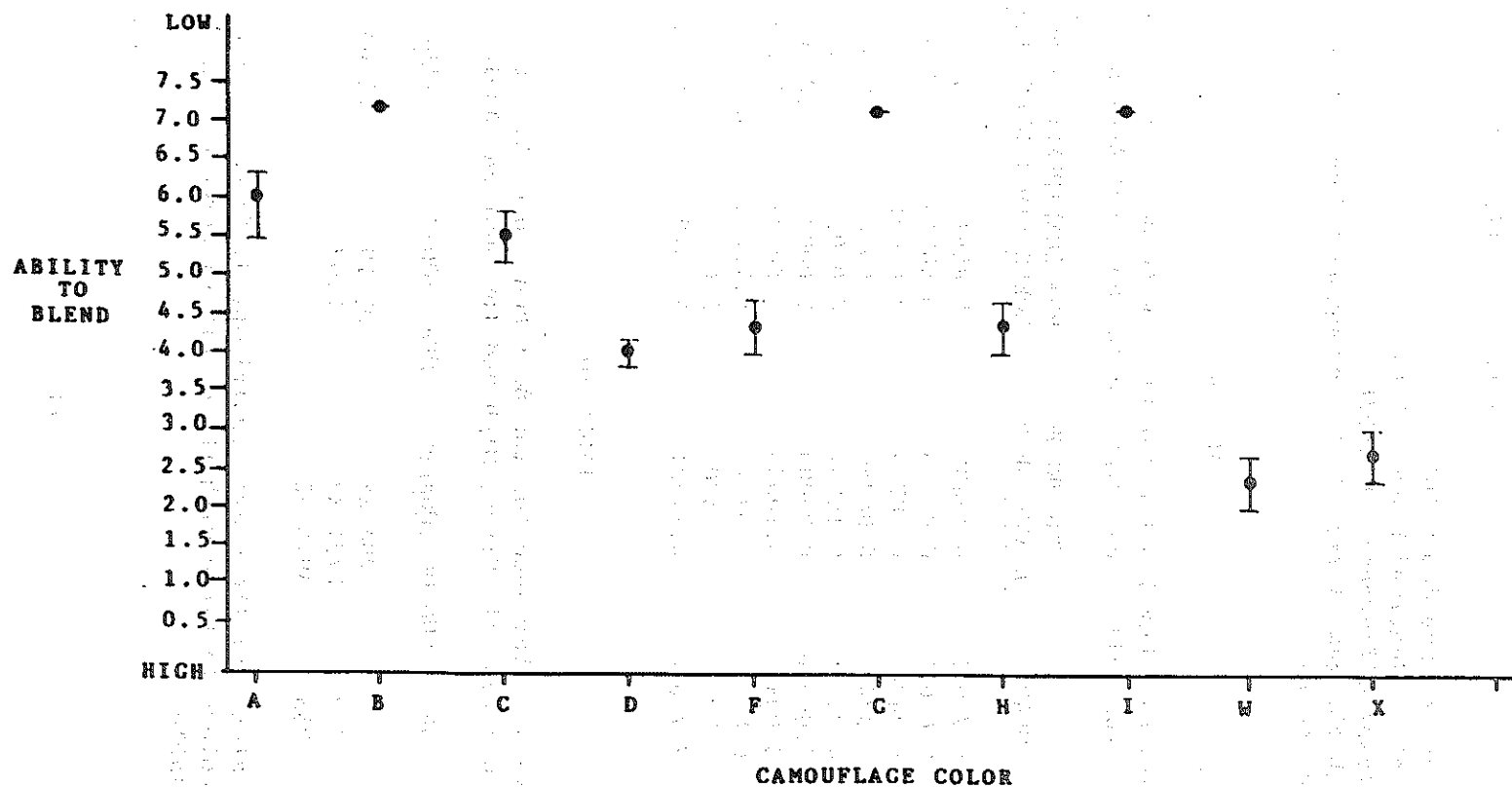


Figure 2 - Camouflage Color Ability to Blend CUCV Truck with Desert Background, Averaged Across All Nine Sites, Low Sunlight Angle

TABLE 8

Significant Differences Between the Top Four Camouflage Colors
(Blend), Averaged Across All Sites, Low Sunlight Angle

<u>TUKEY GROUPING</u>	<u>MEAN</u>	<u>N</u>	<u>COLORS</u>
A	4.1880	117	H
A	3.8120	117	D
B	2.5043	117	X
B	2.1966	117	W

$\alpha = 0.05$, Degrees of Freedom = 464

Critical Value of Studentized Range = 3.646

Minimum Significant Difference = 0.485762

Color means with the same letter in the grouping column are not significantly different.

TABLE 9

Descriptive Data for CUCV Truck Color Blend with Desert Background,
Averaged Across All Sites, High and low Sunlight Angles

<u>COLOR</u>	<u>N</u>	<u>MEAN</u>	<u>STD ERROR OF COL MEAN</u>	<u>95% CONFIDENCE INTERVAL</u>	
				<u>LOWER LIMIT</u>	<u>UPPER LIMIT</u>
A	234	5.76923	0.14495	5.48513	6.05333
B	234	6.63675	0.07875	6.48240	6.79110
C	234	5.04701	0.10719	4.82691	5.25711
D	234	3.82479	0.08236	3.66336	3.98621
F	234	4.24786	0.10557	4.04091	4.45474
G	234	7.00000	0.00000	7.00000	7.00000
H	234	4.00427	0.10053	3.80724	4.20131
I	234	6.85470	0.04513	6.76624	6.94316
W	234	2.90171	0.13803	2.63117	3.17224
X	234	2.71368	0.11821	2.48199	2.83188

TABLE 10

Analysis of Variance for the Best Four Color Blends,
Averaged Across All Sites, High and low Sunlight Angles

<u>SOURCE</u>	<u>DF</u>	<u>SUM OF SQUARES</u>	<u>MEAN SQUARE</u>	<u>F VALUE</u>	<u>PR>F</u>
COLOR	3	294.58	98.19	33.63	0.0001
ERROR	932	2721.37	2.92		
TOTAL	935	3015.95			

Table 10 indicates that there are significant differences in the ability of the top four colors to blend with the desert background. These differences are shown in Table 11.

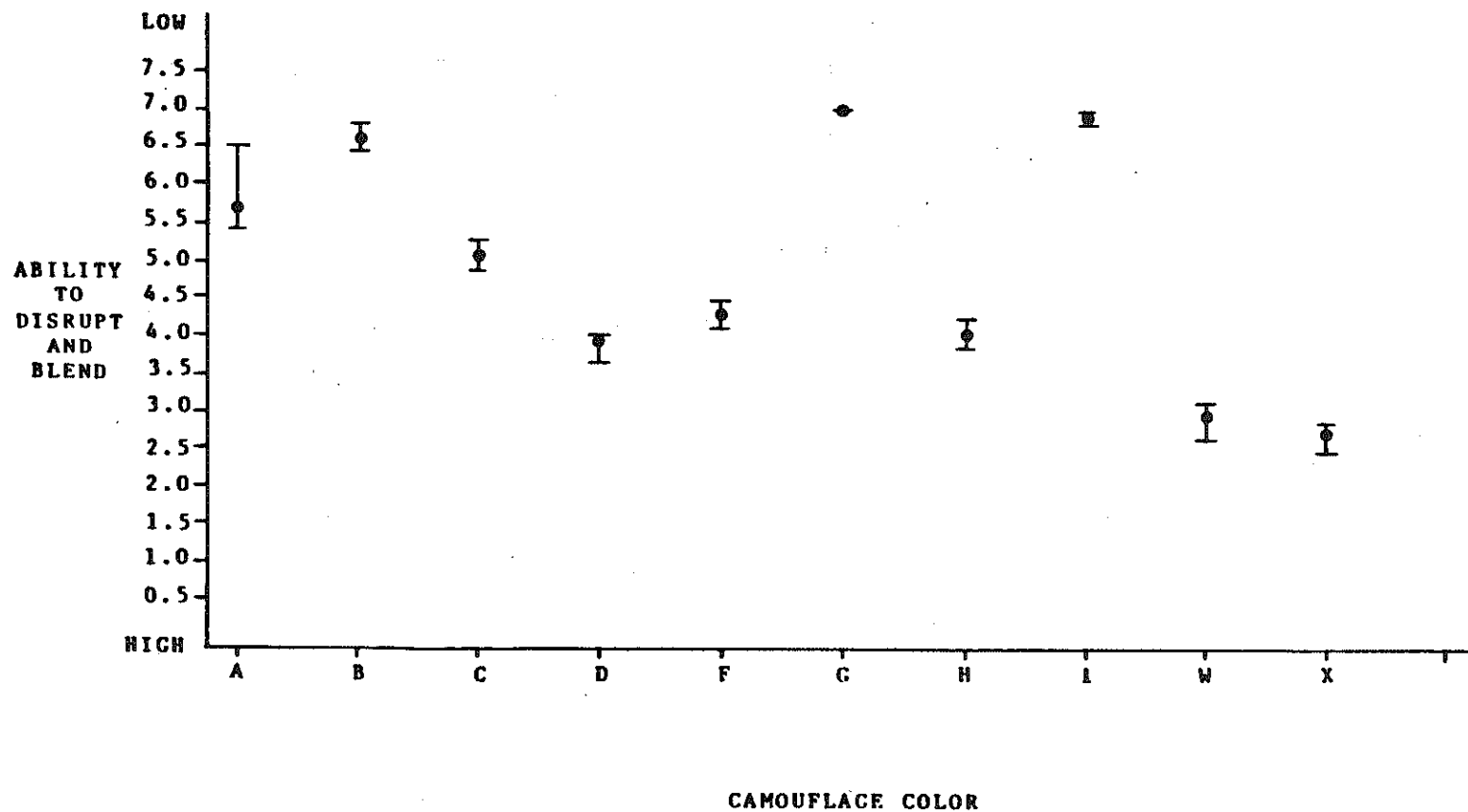


Figure 3 - Camouflage Color Ability to Blend CUCV Truck with Desert Background, Averaged Across All Nine Sites, High and Low Sunlight Angles

TABLE 11

Significant Differences Between the Top Four Camouflage Colors (Blend),
Averaged Across All Sites, High and Low Sunlight Angles

<u>TUKEY GROUPING</u>	<u>MEAN</u>	<u>N</u>	<u>COLORS</u>
A	4.00427	234	H
A	3.82479	234	D
B	2.90171	234	W
B	2.71368	234	X

$\alpha = 0.05$, Degrees of Freedom = 932

Critical Value of Studentized Range = 3.764

Minimum Significant Difference = 0.226501

Color means with the same letter in the grouping column are not significantly different.

4.0 SECTION 4 - Discussion

The purpose of this study was to determine if high and low sunlight angles had a significant effect on the ability of the top four camouflage colors to blend with the desert background. Tables 3-5 and Figure 1 indicate the ability of each of the ten colors evaluated to blend with the desert terrain when averaged across all nine sites for a high sunlight angle. Tables 6-8 and Figure 2 is a repeat of the ability of the ten camouflage paint colors to blend with the terrain, only this time the data was taken under low sunlight conditions. A look at these figures and tables indicates that the conditions of high and low sunlight angles do affect the utility of some of the camouflage colors to blend with the desert terrain. Table 12 shows the best four camouflage colors for each site and when averaged across all nine sites for high and low sunlight angles. For each of the two sunlight angles, the least to most effective colors for blend are read left to right. Thus, there are differences in the best four colors when comparing separately each of the nine sites.

TABLE 12

Summary of the Best Four Color Blends for Each Site and
Across All Sites, High and Low Sunlight Angles

<u>Site</u>	<u>High Sunlight Angle</u>	<u>Low Sunlight Angle</u>
1	B H F A	A H W F
2	C D W X	C D X W
3	C H W X	C D X W
4	C D W X	D A X W
5	H D W X	C D X W
6	H F B A	D H A F
7	X D F H	H F W X
8	C D W X	C D X W
9	D H W X	C D W X
All	D H W X	H D X W

Note that Table 1 shows the colors for each of the alphabetical letters.

For a camouflage color to be effective, it must have camouflage effectiveness across a wide range of sites. It is too costly and time consuming to paint equipment for specific areas unless the resources are to remain in that geographic location for a considerable period of time. Likewise, only the best four camouflage colors should be of interest for this study.

Table 12 shows that the best four camouflage colors to blend with the desert terrain when averaged across all nine sites for high sunlight angle were DHWX, with X the best color and D the worst. The same four colors were also the most effective for the low sunlight angle reading worst to best HDXW. The only difference between the two groups is that the order of X and W and H and D are reversed. For both sunlight angles, colors W and X were better than colors D and H. Therefore, the remaining task is to determine if H and D and W and X differ significantly ($\alpha=0.05$) from each other. Tables 9-11 and Figure 3 indicate the ability of each of the ten colors evaluated to blend with the desert terrain averaged across all nine sites and both high and low sunlight angles. Table 11 indicates that although the colors in color grouping A and B are significantly different ($\alpha=0.05$), there were no significant differences within the groups. Thus, it can be concluded that the reversals of colors H and D and W and X for the high and low sunlight angles are of minor consequence. From a practical field evaluation standpoint, future studies can be conducted using only the high sunlight angle because it represents the longest period of the day.

5.0 SECTION 5 - Summary and Conclusions

A total of ten CUCV vehicles were painted in camouflage colors and viewed by thirteen ground observers at nine desert sites in the United States desert southwest. The colors were divided into two groups of five. The best three colors from each of the two groups were selected on their ability to blend with the desert terrain. The resulting six colors were then ranked on their ability to blend using a six point scale with one being the best and six being the worst. No tie values were allowed and a value of seven was assigned to the colors that did not make the final six. This data was collected for both high and low sunlight angles to determine what effects the lighting conditions had in the rating of the different camouflage colors to blend with the terrain.

Analysis of the data indicated that desert colors W and X were better than H and D for both high and low sunlight angles. The order of W and X and H and D were reversed for the two lighting conditions. Additional statistical analysis revealed that within each color grouping A and B, there were no significant differences ($\alpha=0.05$). The order reversal of H and D and W and X for the two sunlight angle conditions is therefore not important. It is concluded that future field evaluations should involve only one sunlight angle. This will be the high sun angle as it represents a longer period of time for each day.

REFERENCES

1. Anitole, George and Johnson, Ronald, Saudi Arabian National Guard Camouflage Paint Development Program, U.S. Army Mobility Equipment Research and Development Command, Fort Belvoir, Virginia, October 15, 1982
2. Natrella, Mary G., Experimental Statistics, National Bureau of Standards Handbook 91, U.S. Department of Commerce, Washington, D.C., 1966

Weibull Tail Modeling for Estimating Confidence on Quantiles from Censored Samples

Mark Vangel

U.S. Army Materials Technology Laboratory
Watertown, Massachusetts 02172-0001

This paper describes a simple method for estimating lower confidence bounds on quantiles from a Weibull tail model.

A two step procedure is proposed for estimating the 100q% lower confidence bound for the pth quantile of a Weibull sample of size n. Parameter estimates are first obtained for a Weibull model fit to the lower tail values. The inverse of the estimated CDF is then evaluated at the (1-q)th quantile of the beta distribution with parameters $n(1-p)$ and $np+1$.

This method is proposed as a simple alternative to Lawless' elaborate conditional procedure specifically for determining 'B-Basis' values. The B-Basis value is defined to be the quantile corresponding to the lower 95% confidence bound on 90% reliability. This value is used by the aircraft industry to determine the acceptability of composite materials. Composite material failure data is often multimodal, and lower tail modeling is expected to circumvent this difficulty.

A preliminary Monte Carlo study indicates that the proposed method compares favorably with the Lawless procedure for obtaining B-Basis values.

1. Introduction:

When assessing the strength of composite materials for aircraft applications, an important criterion is the material basis property, defined as the 95% lower confidence bound on the stress at which the material fails with 10% probability.

To be useful for this application, a lower confidence bound (LCB) estimator must be able to contend with the primary problems of composite failure data analysis; that is, small samples (≤ 30) and multiple failure modes. Because of this multimodality, a parametric model often cannot be fit to an entire sample, and the standard nonparametric approach (e.g. Conover, 1980), based on the sample order statistics, usually yields very conservative results. In order to get a useful estimate of the basis property in this case, recent work suggests modeling as much of the tail as possible, and considering the rest of the sample as Type II censored (Breiman, Stone, and Gins, 1981). This paper develops a simple approximate method based on such a tail model for estimating confidence bounds on Weibull quantiles, which is particularly useful for estimating material basis properties from small samples.

2. Review of Exact Methods

Exact methods for inference on the parameters of the (two parameter) Weibull distribution

$$F(x) = 1 - e^{-(x/\beta)^\alpha}$$

are ultimately based on the pivotal random variables for the maximum likelihood estimators (MLE's). These pivotals are (Thoman, Bain, and Antle, 1969):

$$Z_1 = \hat{\alpha}/\alpha$$

for the shape parameter (α) and

$$Z_2 = \hat{\alpha} \ln(\hat{\beta}/\beta)$$

for the scale parameter (β). That is, Z_1 and Z_2 have distributions which depend only on the sample size and on the censoring configuration, not on the population parameters. The distributions of these pivots cannot be written down in closed form, but may be easily estimated by Monte Carlo. Once the quantiles of the pivots have been tabulated for various sample sizes, exact confidence intervals for the Weibull MLE's may be obtained.

Confidence on quantiles of the Weibull cumulative distribution function can be calculated from the pivotal for the p th quantile X_p , $0 < p < 1$, which is (Thoman, Bain, and Antle, 1971)

$$Z_p = Z_2 - \ln(-\ln(1-p))Z_1$$

Of course, the quantiles of this pivotal must once again be determined by Monte Carlo. The tables published in the original paper are not always accurate. Corrected tables are available (e.g. Neal and Spiridigliozzi, 1983).

For censored data, it is necessary to tabulate Z_p for censoring situation as well as sample size. Partial tables are available (Billman, Antle, and Bain, 1972), but any reasonably complete tabulation would be unwieldy.

Lawless (1979) demonstrated that although the distribution of Z_p is intractable, the pivotal of the quantile conditioned on the ancillary statistics (statistics whose distribution does not depend on the population parameters) may be found in closed form. With the aid of a computer, a conditional confidence interval for the quantile can then be obtained without resort to Monte Carlo. This conditional interval probably does not differ very much from the unconditional interval (Lawless, 1973).

The Lawless method provides exact conditional intervals for confidence on the parameters and quantiles of any continuous location-scale family, as long as the parameter estimators are equivariant. Equivariant estimators of a location parameter u and a scale parameter b are functions of the sample $\underline{x} = (x_1, \dots, x_n)$ such that for any c_1 and any $c_2 > 0$.

$$\begin{aligned}u(c_1 \underline{x} + c_2) &= c_1 u(\underline{x}) + c_2 \\b(c_1 \underline{x} + c_2) &= c_1 b(\underline{x})\end{aligned}$$

In particular, MLE's are equivariant estimators. A detailed development of the conditional procedure may be found in Lawless (1982).

Since the logarithm of a random variable having the extreme value distribution with location u and scale b ,

$$G(x) = e^{-((x-u)/b)},$$

is Weibull with shape (α) and scale (β) given by

$$\alpha = 1/b \quad \beta = e^u$$

the Lawless procedure applied to the extreme value distribution will yield the desired confidence on the Weibull quantile. This procedure is sketched below for Type II censoring. This outline follows the exposition in Lawless' book (1982).

If the Type II censored sample

$$x_1, x_2, \dots, x_r \quad r \leq n$$

is independently identically distributed $G(x)$, and if \hat{u} and \hat{b} are any equivariant estimators of the extreme value parameters, then:

$$Z_1 = (\hat{u} - u)/\hat{b} \quad Z_2 = \hat{b}/b \quad Z_3 = (\hat{u} - u)/b$$

$$Z_p = Z_1 - \ln(-\ln(1-p))/Z_2$$

are all pivotal statistics; with Z_p pivotal for the p th quantile of $G(x)$. Also, the statistics:

$$\underline{a} = \left\{ (x_i - \hat{u})/\hat{b}; i=1, \dots, r \right\}$$

form a complete set of ancillary statistics of which any $r-2$ are

functionally independent.

Let the corresponding ordered extreme value sample be

$$y_1 \leq y_2 \leq \dots \leq y_r$$

The conditional pdf of Z_2 given \underline{a} is of the form

$$h_2(z|\underline{a}) = \frac{k(\underline{a}, r, n) e^{(z-1)\sum_{i=1}^r a_i}}{((\sum_{i=1}^r e^{a_i z})/r)^r}$$

where K is a constant given \underline{a} , r , and n . The constant is determined by numerically integrating the density $h_2(z|\underline{a})$. Finally, the conditional distribution of Z_p given \underline{a} is

$$P(Z_p \leq t|\underline{a}) = \int_0^{\infty} h_2(z|\underline{a}) I(r, e^{w_p + tz} \sum_{i=1}^r e^{a_i z}) dz$$

where

$$\sum_{i=1}^r w_i^* = \sum_{i=1}^r w_i + (n-r)w_r$$

$$w_p = \ln(-\ln(1-p))$$

and $I(r, s)$ is the incomplete gamma function

$$I(r, s) = \frac{1}{\Gamma(r)} \int_0^s u^{r-1} e^{-u} du$$

The Lawless method may be used to calculate exact conditional confidence intervals or bounds for Weibull quantiles without the need for tables. The primary disadvantage of this procedure is its complexity. The numerical integration is not trivial, particularly when r is large. It is the aim of this paper to present a very simple approximate method for obtaining intervals which are often close to the Lawless results.

3. An Approximate Method for Estimating the LCB of a Quantile

Let $y_1 \leq \dots \leq y_r$ be the r smallest order statistics from a continuous distribution $F(\cdot)$. Let

$$x_p = F^{-1}(p)$$

be the p th quantile of $F(x)$, and let L_p be an estimated $100\gamma\%$ LCB for X_p . Assume initially that $p = j/n$ for some integer j so that y_j estimates X_p .

Using y_j as an estimator for X_p , one obtains the following approximation

$$\gamma \approx P(\hat{L}_p \leq x_p) = P(F(\hat{L}_p) \leq F(x_p)) \approx 1 - P(F(y_j) \geq F(\hat{L}_p)).$$

But $F(y_j)$ has the beta distribution

$$1-\gamma = \text{Beta}(u_\gamma | j, n-j+1) = \frac{\Gamma(j)\Gamma(n-j+1)}{\Gamma(n+1)} \int_0^u t^{j-1}(1-t)^{n-j} dt.$$

The approximate LCB is then

$$\hat{L}_p = F^{-1}(u_\gamma).$$

If $j/n = p$ for integer j , let u_γ be the $100(1-\gamma)\%$ percentile from the Beta $(u; pn, (1-p)n+1)$ distribution.

For the Weibull case

$$\hat{L}_p = \hat{F}^{-1}(u_\gamma) = \hat{\beta} \ln(1/(1-u_\gamma))^{1/\hat{\alpha}}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the MLE's. This estimator is identical to

Thoman, Bain, and Antle's estimator, except the quantile of Z_p for appropriate n and r is replaced with a quantile from a beta distribution.

4. Interpretation of the LCB Estimator as an approximation to the quantile pivotal

Following Thoman, Bain, and Antle (1971), let the distribution of Z_p be $G(Z)$ and

$$P(Z_p \leq z_\gamma) = G(z_\gamma) = \gamma.$$

This implies that

$$(*) \quad P(\hat{\beta} e^{-z_\gamma / \hat{\alpha}} \leq \beta(-\ln(1-p))^{1/\alpha}) = \gamma$$

it is because of (*) that Z_p is pivotal for X_p . The new estimator yields an approximate relation of the same form as (*).

$$P(\hat{\beta}(-\ln(1-u_\gamma))^{1/\hat{\alpha}} \leq \beta(-\ln(1-p))^{1/\alpha}) \approx \gamma$$

For this to be an approximation, of course, the left hand sides of the inequalities should be nearly equal

$$\beta(-\ln(1-u_\gamma))^{1/\hat{\alpha}} \approx \hat{\beta} e^{-z_\gamma / \hat{\alpha}}$$

or, equivalently,

$$z_\gamma \approx \tilde{z}_\gamma = -\ln(-\ln(1-u_\gamma)).$$

For the approximation to be useful, the random variable should have a distribution close to that of the pivotal Z_p in the vicinity of the quantiles of interest. Since \tilde{Z} is a simple transformation of a beta random variable, if

$$u(\tilde{z}) = 1 - e^{-e^{-\tilde{z}}}$$

then the density of z is

$$f(u(\tilde{z})) = \frac{\Gamma(n+1)\Gamma(u-1)}{\Gamma(r)\Gamma(n-r+1)} u^{r-1} (1-u)^{n-r}.$$

To graphically illustrate the agreement between the pivotal density and the density of \tilde{z} , several simulations were performed, (Figure 1, a-d). The values of p and γ were set at .1 and .95 respectively, since the 95% lower confidence bound on 10% probability of failure is the case of primary interest in aircraft design. The sample sizes were kept small - reflecting the expected range of sample sizes of composite failure data, $n = 10, 20, 30, 40$, and 50. For each sample size, the upper two thirds of the data was Type II censored: $r = 6, 9, 12$, and 15. For \tilde{z} , the exact density is plotted. For the pivotal, the density is estimated using a four parameter generalization of Tukey's lambda distribution (Ramberg, et.al., 1979) applied to 2,500 Monte Carlo replicates for each case. The agreement between the densities appears to be quite good, as long as one bears in mind that for intervals with reasonable confidence, one need only be concerned with the validity of the approximation in the tails.

5. Comparison with the Lawless Method

A simulation was performed to directly compare the Lawless procedure with the approximation presented in this paper. Because of the computational effort required for the Lawless integration, the scope of this study was necessarily modest. However, useful results were obtained despite the restriction to 10 replicates per case. It was decided to fix $p = .1$ and $\gamma = .95$ as in the previous section. Also, the sample size was fixed at 30, since this is typical for composite material failure data in aircraft industry testing. Lower confidence bound estimates were obtained for pseudo-random Weibull samples with shape parameters in the range 2 to 100 and Type II censoring of 90% to 0% ($r = 3, 6, 9, \dots, 30$). The average percent differences in the results are presented in Figure 2a. Note that for $r = 9$, there is amazing agreement between

the two methods. This could have been anticipated from the close agreement at the 95th percentiles of Z and Z for this case (Figure 1d).

For the approximate estimator, the dependence of the simulation results on the Weibull shape parameter may be completely removed by transforming the estimator L_p to its pivotal:

$$\alpha \ln(\hat{L}_p / \beta)$$

This transformation was applied to both the Lawless results and the approximation results. The percent difference between the methods for the transformed data showed no dependence on r , so these were averaged over all the data providing a measure of percent difference vs. r based on 150 replicates per r value. (Figure 2b). Positive percent difference is defined here to mean that the Lawless bound was greater than the approximate bound. For $9 \leq r \leq 30$, the approximation yields a conservative result. It is reassuring that potentially dangerous nonconservative estimates only occur for very small values of r .

6. Examples

As examples, the approximate method was applied to three extreme value data sets from the literature (Figure 3 and 4). In all of these cases, either the approximation gives a result very close to that obtained via the conditional procedure, or the approximation provides a result which is more conservative.

These examples, of course, cannot by themselves validate the proposed method. They are intended rather to highlight the ease with which one may arrive at reasonable results, making use of a computer only to obtain MLE's of the parameters and, possibly, the quantiles of the relevant beta distribution.

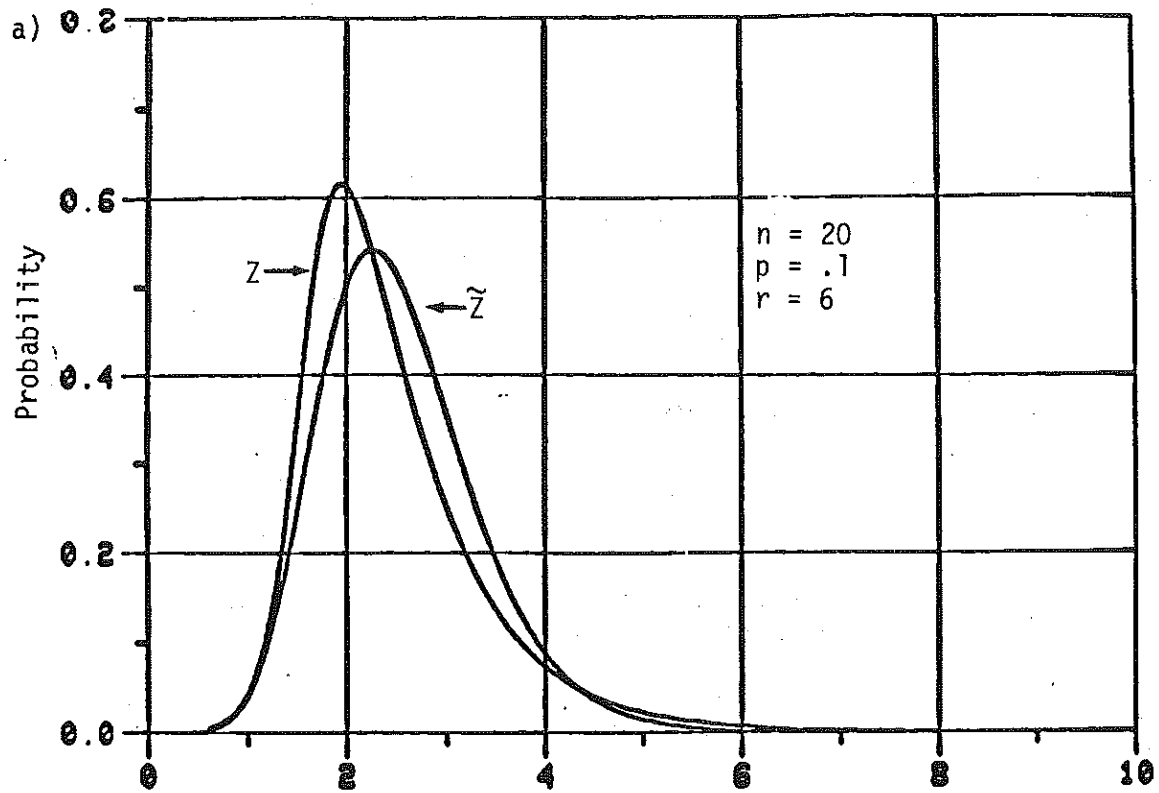
7. Conclusion

The proposed method is attractive as an alternative to the Lawless procedure. The Lawless method is computationally complex, whereas the new method is very easy to apply. Unfortunately, while the Lawless method may be justified theoretically, the proposed method as yet has no firm theoretical basis. The interpretation of the new method as an approximation to the pivotal is interesting, but by itself it cannot provide this foundation. The natural question of how good this approximation is in general cannot be answered because the pivotal distribution can only be obtained by simulation. For the cases considered, namely 95% LCB on 10% point from samples of 10 through 50, however, the approximation is good. Also, the method has been demonstrated to give results for a sample size of 30, which are generally either close to or more conservative than the Lawless results. To validate the procedure, either an extensive Monte Carlo study or a deeper theoretical investigation must be performed. Both of these approaches will be considered in the near future.

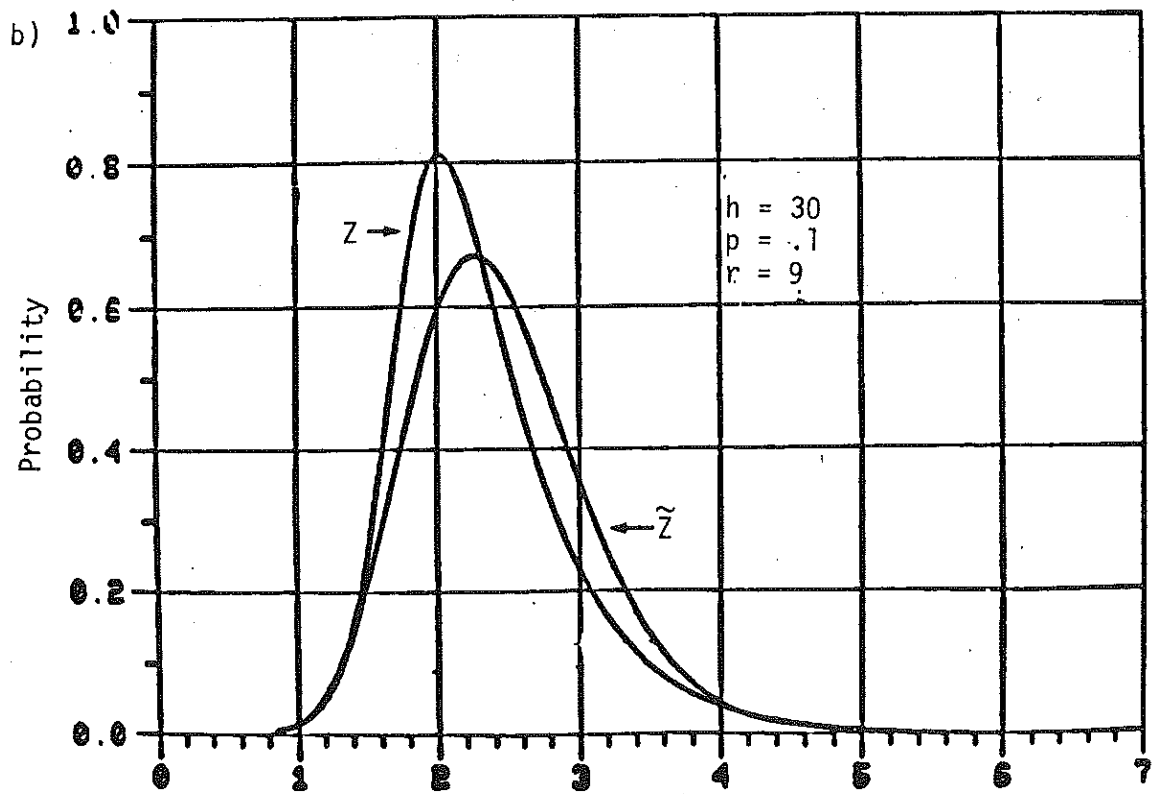
References:

- Billman, B. R., C. E. Antle, and L. J. Bain, (1972), "Statistical Inference from Censored Weibull Samples", Technometrics, 14, 831-839.
- Breiman, L., C. Stone, and J. Gins, (1981), "Further Developments of New Methods for Estimating Tail Probabilities and Extreme Value Distributions", Technical Report #TSD-PH-A243-1, Technology Services Corporation, Santa Monica, CA.
- Conover, W. J., (1980), Practical Nonparametric Statistics, Second Edition, New York: John Wiley.
- Lawless, J. F., (1973), "Conditional vs. Unconditional Confidence Intervals for the Parameters of the Weibull Distribution", Journal of the American Statistical Association, 68, 665-669.
- _____, (1975), "Construction of Tolerance Bounds for the Extreme Value and Weibull Distributions", Technometrics, 17, 255-261.
- _____, (1982), Statistical Models and Methods for Lifetime Data, New York: John Wiley.
- Neal, D., and L. Spiridigliozzi, (1983), "An Efficient Method for Determining the 'A' and 'B' Design Allowable", Proceedings of the 28th Conference on the Design of Experiments in Army Research Development and Testing, 199-235.
- Ramberg, J. S., E. S. Dudewicz, P. R. Tadikamalla, and E. F. Mykytka (1979), "A Probability Distribution and its Uses in Fitting Data", Technometrics, 21, 201-214.
- Thoman, D. R., L. S. Bain, and C. E. Antle, (1969). "Inference on the Parameters of the Weibull Distribution", Technometrics, 11, 445-459.
- _____, (1970), "Maximum Likelihood Exact Confidence Intervals for Reliability and Tolerance Limits on the Weibull Distribution", Technometrics, 12, 363-371.

Comparison of Pivotal (Z) with Approximation (\tilde{Z})



$$Z_{.95} = 4.12 \quad \tilde{Z}_{.95} = 4.00$$



$$Z_{.95} = 3.57 \quad \tilde{Z}_{.95} = 3.57 \quad 224$$

FIGURE 1 (continued)

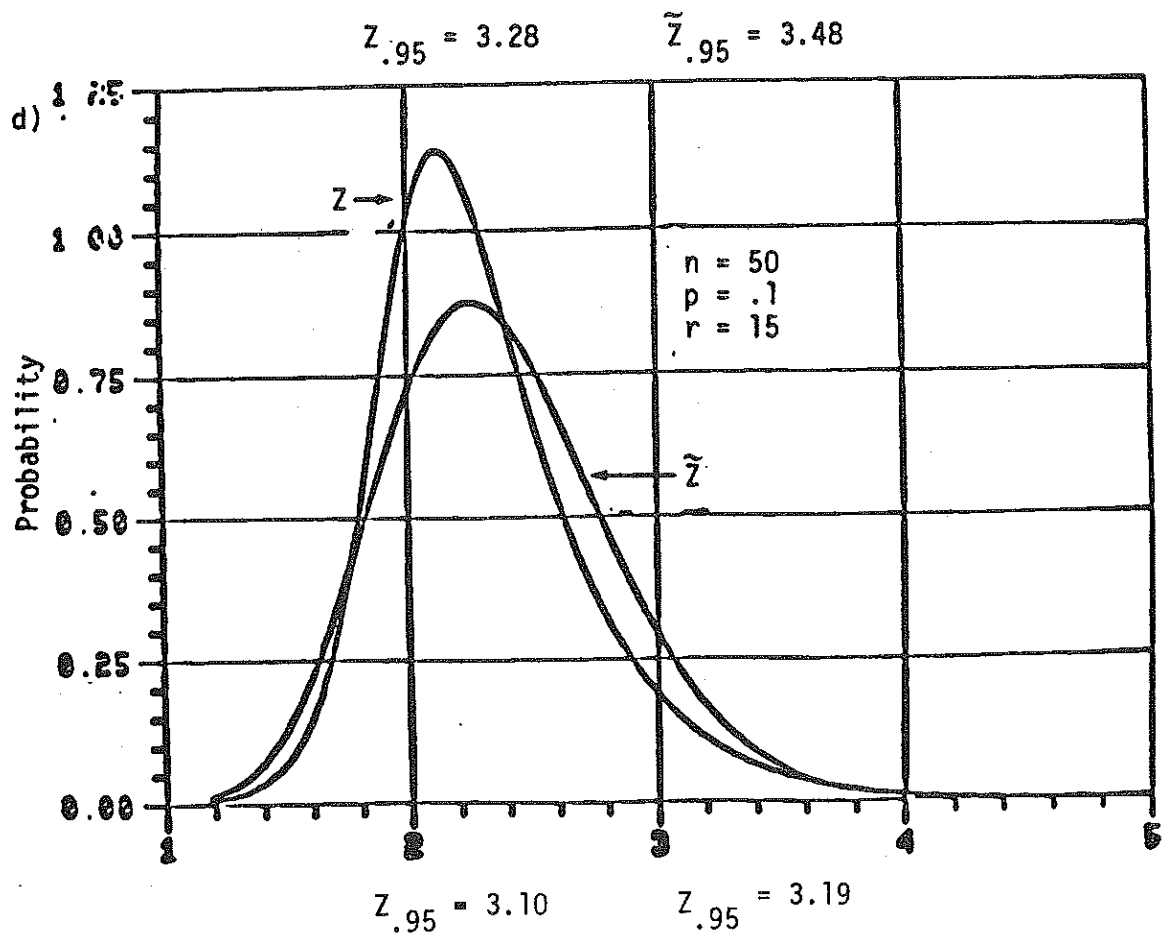
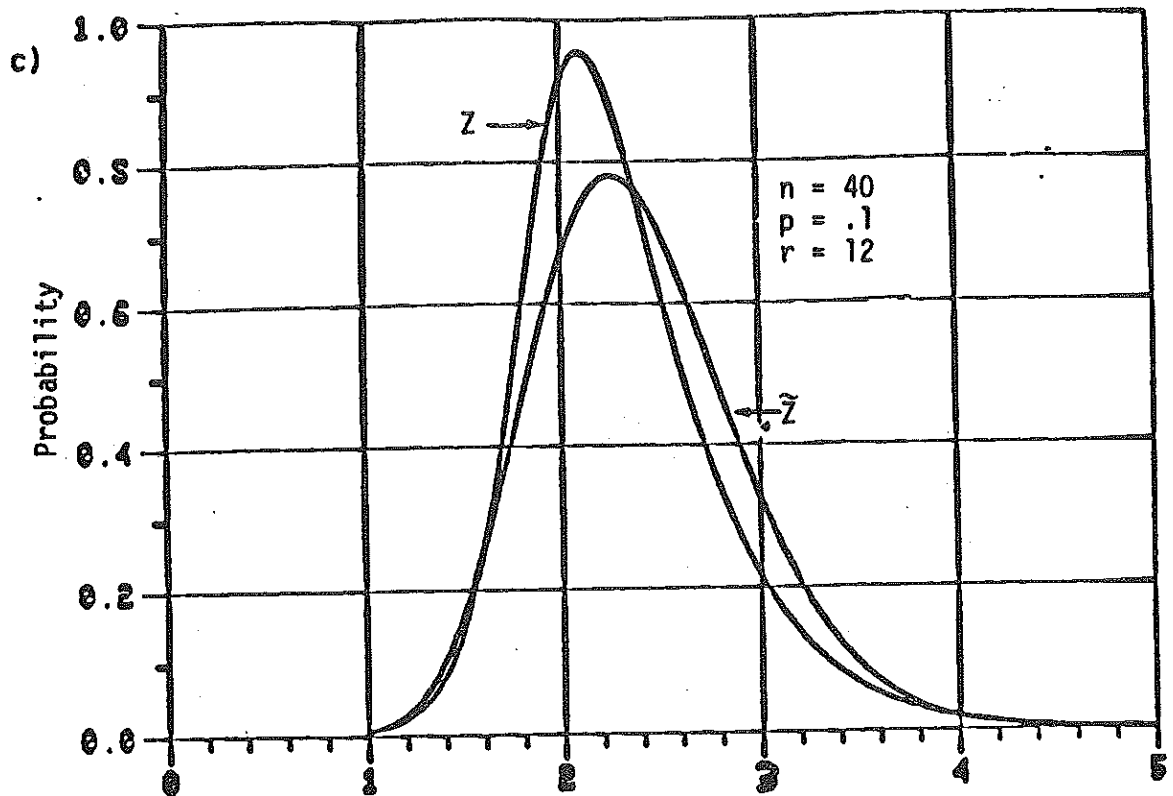


FIGURE 2
MONTE CARLO COMPARISON OF
APPROXIMATE METHOD WITH LAWLESS METHOD

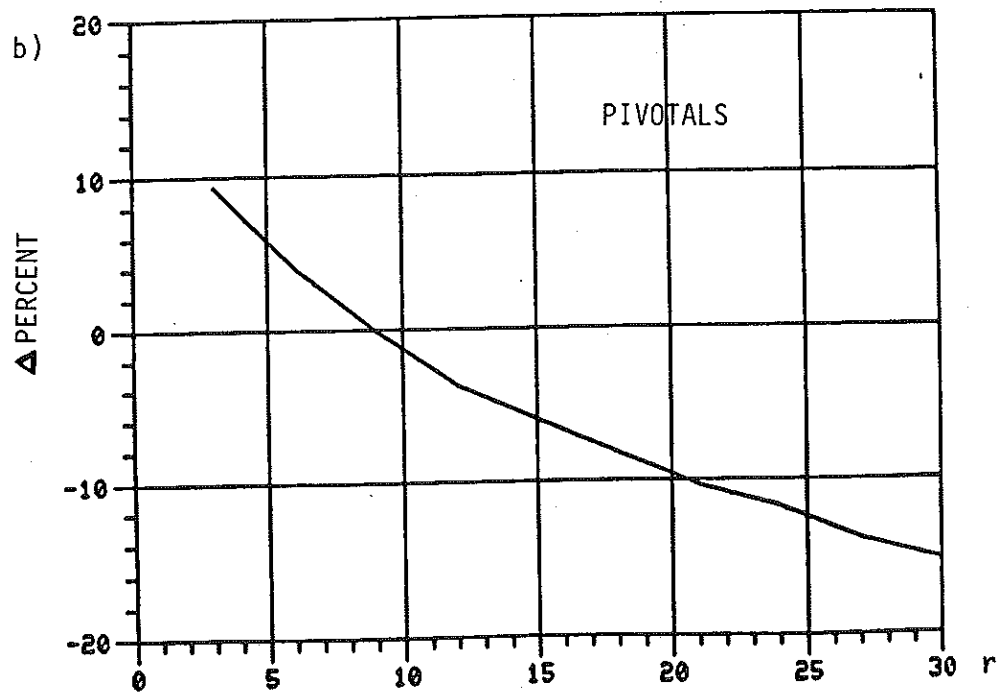
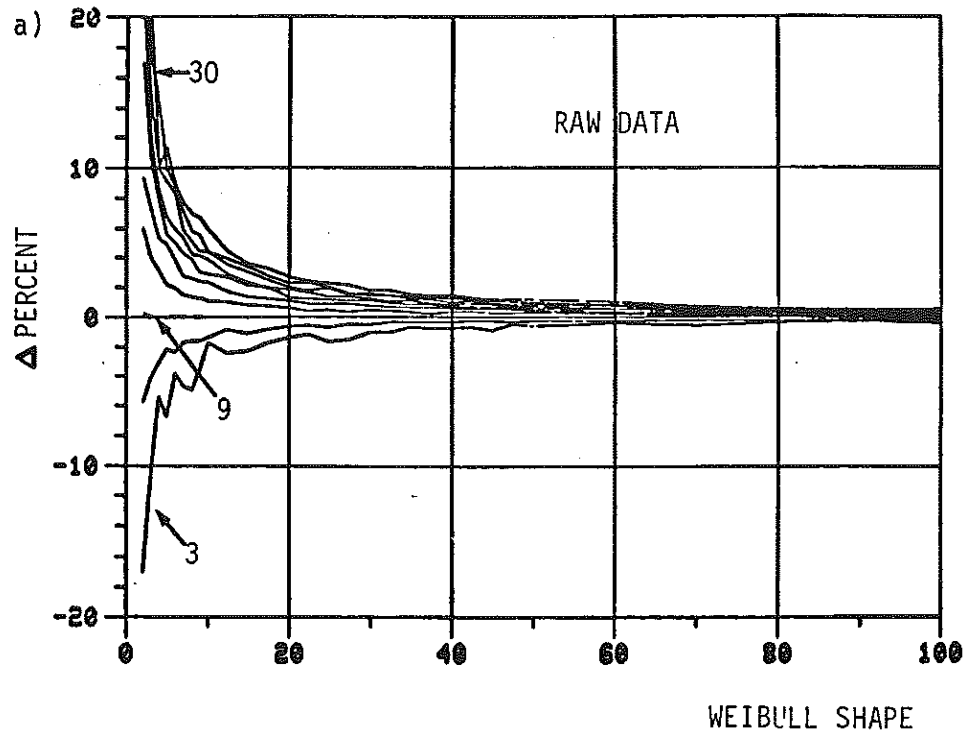


FIGURE 3

EXAMPLE: LAWLESS (1982), p.156

Type II censored extreme value sample

$$n = 20 \quad r = 10$$

Estimate 90% confidence interval for $X_{.1}$

$$\hat{u} = -.122 \quad \hat{b} = .907 \quad \hat{a} = .931 / \hat{b} = 1.026^*$$

$$\hat{\beta} = e^{\hat{u}} = .8852$$

$$\int_0^{s_1} \text{Beta}(t; 1, 19) dt = \int_{s_1}^1 \text{Beta}(t; 1, 19) dt = .05$$

$$\ln [-\hat{\beta} \ln(1-s_1)^{1/\hat{a}}] = -4.03, \text{ Lawless} = -3.74$$

$$\ln [-\hat{\beta} \ln(1-s_2)^{1/\hat{a}}] = -1.50, \text{ Lawless} = -1.49$$

* Unbiased MLE (Thoman, Bain, and Antle, 1969)

FIGURE 4

EXAMPLE: LAWLESS (1975), p. 255

$n = 40$ $r = 28$ $-2.982, -2.849, . . . , .245, .296$

Pseudo random sample from extreme value distribution with $u = 0, b = 1$

$$\hat{u} = .1563 \quad \hat{b} = .9104$$

$$\hat{a} = .966 / \hat{b} = 1.061$$

$$\hat{\beta} = e^{\hat{u}} = 1.169$$

Lower 95% confidence on

	$X_{.1}$	$X_{.05}$
Lawless	-2.71	-3.61
Approximation	-2.99	-3.62

THE LINDSTROM-MADDEN METHOD FOR SERIES SYSTEMS
WITH REPEATED COMPONENTS

Andrew P. Soms

Department of Mathematical Sciences
University of Wisconsin-Milwaukee, Milwaukee, WI 53201
Mathematics Research Center
University of Wisconsin-Madison, 610 Walnut Street
Madison, WI 53705

Key Words and Phrases: confidence limits; reliability

ABSTRACT

The Lindstrom-Madden method of computing lower confidence limits for series systems with unlike components is extended to series systems with repeated components utilizing the results of Harris and Soms (1983). An exact solution is given for no failures and key test results, together with an approximation for the general case. Numerical examples are also provided.

1. INTRODUCTION AND SUMMARY

A problem of substantial importance to practitioners in reliability is the statistical estimation of the reliability of a series system of stochastically independent components when some components are repeated, using experimental data collected on the individual components. In the situations discussed in this paper, the component data consist of a sequence of Bernoulli trials. Thus, for component i , $i = 1, 2, \dots, k$, the data is the pair (n_i, Y_i) , where n_i is the number of trials and Y_i is the

number of observations for which the component functions.

Y_1, Y_2, \dots, Y_k are assumed to be mutually independent random variables. We assume that there are γ_i components of type i , $1 \leq i \leq k$. Then the parameter of interest is

$h(p_1, p_2, \dots, p_k) = h(\tilde{p})$, the reliability of the system, where

$$h(\tilde{p}) = \prod_{i=1}^k p_i^{\gamma_i}.$$

More specifically, it is desired to obtain a Buehler (1957) optimal lower $1 - \alpha$ confidence limit on $h(\tilde{p})$.

The case of $\gamma_1 = \gamma_2 = \dots = \gamma_k = 1$ has been treated in Sudakov (1974), Winterbottom (1974), and Harris and Soms (1983).

In Section 2 we summarize the general theory of Harris and Soms (1983) applicable here. In Section 3 the exact solutions to no failures and key test results are given. Lindstrom-Madden type approximations are given in Section 4. Section 5 contains numerical examples.

2. BUEHLER'S METHOD FOR OPTIMAL CONFIDENCE LIMITS

We now specialize the general results of Harris and Soms (1983) on optimal confidence limits for system reliability to a series system with independent and repeated components. As in Section 1, let

$$h(\tilde{p}) = \prod_{i=1}^k p_i^{\gamma_i},$$

$0 \leq p_i \leq 1$, $X_i = n_i - Y_i$, $x_i = n_i - y_i$, $1 \leq i \leq k$,
 $S = \{\tilde{x} | x_i = 0, 1, \dots, n_i, 1 \leq i \leq k\}$ and let $g(\tilde{x}) = (x_1, x_2, \dots, x_k)$ be an ordering function, i.e., for real x_i , $0 \leq x_i \leq n_i$, $g(\tilde{x})$ is non-decreasing in each component. It is often convenient to normalize $g(\tilde{x})$ by letting $g(\tilde{0}) = 1$ and $g(\tilde{n}) = 0$. With such a normalization, $g(\tilde{x})$ is often selected to be a point estimator of $h(\tilde{p})$. Also let $R = \{r_1, r_2, \dots, r_s, s \geq 2\}$ be the range set of $g(\tilde{x})$. With no loss of generality we order R so that

$r_1 > r_2 > \dots > r_s$ and let $A_i = \{\tilde{x} | g(\tilde{x}) = r_i, \tilde{x} \in S, i = 1, 2, \dots, s\}$. The sets A_i constitute a partition of S induced by $g(\tilde{x})$. We assume throughout that the data is distributed by

$$\begin{aligned} f(\tilde{x}; \tilde{p}) &= P_{\tilde{p}}(\tilde{X} = \tilde{x}) = \prod_{i=1}^k \binom{n_i}{x_i} p_i^{x_i} q_i^{n_i - x_i} \\ &= \prod_{i=1}^k \binom{n_i}{y_i} p_i^{y_i} q_i^{n_i - y_i}, \end{aligned} \quad (2.1)$$

where $q_i = 1 - p_i$, $i = 1, 2, \dots, k$. With no loss of generality, we assume $n_1 < n_2 < \dots < n_k$.

From these definitions, it follows that

$$P_{\tilde{p}}\left\{X \in \bigcup_{i=1}^j A_i\right\} = P_{\tilde{p}}\{g(\tilde{X}) > r_j\}. \quad (2.2)$$

From (2.1) and (2.2), we have

$$P_{\tilde{p}}\{g(\tilde{X}) > r_j\} = \sum_{i_1=0}^{u_1} \sum_{i_2=0}^{u_2} \dots \sum_{i_k=0}^{u_k} f(\tilde{i}; \tilde{p}), \quad (2.3)$$

where $\tilde{i} = (i_1, i_2, \dots, i_k)$ and $u_2 = u_2(i_1), \dots, u_k = u_k(i_1, i_2, \dots, i_{k-1})$ are integers determined by r_j . Equivalently,

$$P_{\tilde{p}}\{g(\tilde{X}) > r_j\} = \sum_{i_1=0}^{[t_1]} \sum_{i_2=0}^{[t_2]} \dots \sum_{i_k=0}^{[t_k]} f(\tilde{i}; \tilde{p}), \quad (2.4)$$

where $t_2 = t_2(i_1), \dots, t_k = t_k(i_1, i_2, \dots, i_{k-1})$, with $t_1 = \sup\{t | 0 \leq t \leq n_1 \text{ and } g(t, 0, 0, \dots, 0) > r_j\}$ and $t_\ell(i_1, i_2, \dots, i_{\ell-1}) = \sup\{t | 0 \leq t \leq n_\ell \text{ and } g(i_1, i_2, \dots, i_{\ell-1}, t, 0, \dots, 0) > r_j\}$, $\ell = 2, 3, \dots, k$.

We now introduce the notion of Buehler optimal confidence limits. Let $g(x) = r_j$. Then define

$$a_{g(\tilde{x})} = \inf\{h(\tilde{p}) | P_{\tilde{p}}\{\tilde{i} | g(\tilde{i}) > g(\tilde{x})\} > \alpha\}. \quad (2.5)$$

Equivalently, by (2.2), we can also write

$$a_{g(\tilde{x})} = \inf_{\tilde{p}} \{ h(\tilde{p}) \mid P_{\tilde{p}} \{ x \in \bigcup_{i=1}^j A_i \} > \alpha \} . \quad (2.6)$$

Then we have, from Harris and Soms (1983),

Theorem 2.1. $a_{g(\tilde{x})}$ is a $1 - \alpha$ lower confidence limit for $h(\tilde{p})$. If $b_{g(\tilde{x})}$ is any other $1 - \alpha$ lower confidence limit for $h(\tilde{p})$ with $b_{r_1} > b_{r_2} > \dots > b_{r_j}$, then $b_{g(\tilde{x})} \leq a_{g(\tilde{x})}$ for all $\tilde{x} \in S$.

Two possible choices of $g(\tilde{x})$ are

$$g(\tilde{x}) = \prod_{i=1}^k ((n_i - x_i)/n_i)^{\gamma_i} , \quad (2.7)$$

or

$$g(\tilde{x}) = \prod_{i=1}^k \prod_{j=0}^{\gamma_i-1} \left(\frac{n_i - x_i - j}{n_i - j} \right) . \quad (2.8)$$

Both reduce to the generally used $g(\tilde{x})$ for series systems with independent components when $\gamma_1 = \gamma_2 = \dots = \gamma_k = 1$, i.e.,

$$g(\tilde{x}) = \prod_{i=1}^k (n_i - x_i)/n_i .$$

Since (2.7) is the maximum likelihood estimator of $h(\tilde{p})$ we will use it here and from now on it will be understood that $g(\tilde{x})$ is given by (2.7). With this choice of $g(\tilde{x})$, we assume from now on that $0 \leq x_i < n_i$, $i = 1, 2, \dots, k$, since $a_{g(\tilde{x})} = 0$ if some $x_i = n_i$. With this assumption, the t_i in (2.4) are given by

$$t_1 = n_1 - \left(\prod_{i=1}^k (n_i - x_i)^{\gamma_i} / \prod_{i=2}^k n_i^{\gamma_i} \right)^{1/\gamma_1} \quad (2.9)$$

and

$$t_\ell = n_1 - \left(\prod_{i=1}^k (n_i - x_i)^{\gamma_i} \right) / \left(\prod_{s=1}^{\ell-1} (n_s - i_s)^{\gamma_s} \prod_{i=\ell+1}^k n_i^{\gamma_i} \right)^{1/\gamma_\ell}, \quad (2.10)$$

$$\ell = 2, \dots, k, \text{ with } \prod_{i=k+1}^k n_i^{\gamma_i} = 1.$$

For the purpose of simplifying the calculation of $a_{g(\tilde{x})}$ in special cases it is necessary to state additional results from Harris and Soms (1983).

Theorem 2.2. Let $g(\tilde{x}) = r_j$ and let

$$f^*(\tilde{x}; a) = \sup_{h(p)=a} P\{g(\tilde{X}) > r_j\}, \quad 0 < a < 1. \quad (2.11)$$

Then

$$\inf_{0 < a < 1} f^*(\tilde{x}; a) = 0, \quad \sup_{0 < a < 1} f^*(\tilde{x}; a) = 1$$

and $f^*(\tilde{x}; a)$ is strictly increasing in a .

Theorem 2.3. $f^*(\tilde{x}; a) = \alpha$ has exactly one solution a_α in a and $a_\alpha = a_{g(\tilde{x})}$.

3. EXACT SOLUTIONS FOR ZERO FAILURES AND KEY TEST RESULTS

We first assume that $\tilde{x} = (0, 0, \dots, 0) = \tilde{0}$ and use Theorem 2.3 to obtain $a_{g(\tilde{0})}$.

Theorem 3.1. If $\tilde{x} = \tilde{0}$, then

$$f^*(\tilde{0}; a) = \sup_{\prod_{i=1}^k p_i^{\gamma_i} = a} \prod_{i=1}^k p_i^{n_i} = a^{n_j/\gamma_j}, \quad (3.1)$$

where $n_j/\gamma_j = \min_{1 \leq i \leq k} n_i/\gamma_i$ and

$$a_{g(\tilde{0})} = \alpha^{n_j/\gamma_j}. \quad (3.2)$$

Proof.

$$\prod_{i=1}^k p_i^{n_i} = \left(\prod_{i=1}^k p_i^{\gamma_i} \right)^{n_j/\gamma_j} \prod_{\substack{i=1 \\ i \neq j}}^k p_i^{(n_i \gamma_j - n_j \gamma_i)/\gamma_j}$$

$$\leq a^{n_j/\gamma_j},$$

since $n_i \gamma_j - n_j \gamma_i > 0$ is equivalent to $n_i/\gamma_i > n_j/\gamma_j$, which is true, and therefore $\prod_{\substack{i=1 \\ i \neq j}}^k p_i^{(n_i \gamma_j - n_j \gamma_i)/\gamma_j} < 1$. (3.1) follows by

noting that the choice $p_j = a^{1/\gamma_j}$, $p_i = 1$, $i \neq j$, gives $\prod_{i=1}^k p_i^{n_i} = a^{n_j/\gamma_j}$. Then, using Theorem 2.3, we obtain (3.2),

which reduces to the known series result if

$$\gamma_1 = \gamma_2 = \dots = \gamma_k = 1.$$

We now turn to analogues of key test results (see, e.g., Winterbottom (1974) and Harris and Soms (1983)). We define a key test result if $\gamma_1 = \max_{1 \leq i \leq k} \gamma_i$ (recall that $n_1 = \min_{1 \leq i \leq k} n_i$) and $\tilde{x} = (x_1, 0, \dots, 0)$.

Theorem 3.2. If \tilde{x} is a key test result and

$$\begin{aligned} \left\{ \tilde{z} \mid \prod_{i=1}^k (n_i - z_i)^{\gamma_i} > \prod_{i=1}^k (n_i - x_i)^{\gamma_i} \right\} &= \left\{ \tilde{z} \mid \sum_{i=1}^k (n_i - z_i) \right. \\ &> \left. \sum_{i=1}^k (n_i - x_i) \right\}, \end{aligned} \quad (3.3)$$

then

$$f^*(\tilde{x}; a) = I_{1/\gamma_1}^{n_1 - x_1, x_1 + 1} (a), \quad (3.4)$$

where $I_x(a, b)$ is the incomplete beta function. Let b_α denote the solution in b of

$$\alpha = I_b(n_1 - x_1, x_1 + 1).$$

Then $a_{g(\tilde{x})} = b_{\alpha}^{\gamma_1}$. Note that b_{α} is the usual $1 - \alpha$ lower confidence limit on p , given x_1 failures in n_1 trials.

Proof. Without loss of generality we can assume that

$n_1 = n_2 = \dots = n_k$, for otherwise we can write (2.4) as

$$\begin{aligned}
 P_p\{g(\tilde{X}) > r_j\} &= \sum_{i_1=0}^{x_1} \binom{n_1}{i_1} p_1^{n_1-i_1} q_1^{i_1} \sum_{i_1=0}^{x_1-i_1} \binom{n_2}{i_2} p_2^{n_2-i_2} q_2^{i_2} \dots \\
 &\quad \sum_{i_{k-1}=0}^{x_1-i_1-i_2-\dots-i_{k-2}} \binom{n_{k-1}}{i_{k-1}} p_{k-1}^{n_{k-1}-i_{k-1}} q_{k-1}^{i_{k-1}} I_{p_k}(n_k - \\
 &\quad (x_1-i_1-i_2-\dots-i_{k-1}), x_1-i_1-i_2-\dots-i_{k-1}+1) \\
 &< \sum_{i_1=0}^{x_1} \binom{n_1}{i_1} p_1^{n_1-i_1} q_1^{i_1} \dots \sum_{i_{k-1}=0}^{x_1-i_1-i_2-\dots-i_{k-2}} \binom{n_{k-1}}{i_{k-1}} p_{k-1}^{n_{k-1}-i_{k-1}} q_{k-1}^{i_{k-1}} I_{p_k}(n_1 - \\
 &\quad (x_1-i_1-i_2-\dots-i_{k-1}), x_1-i_1-i_2-\dots-i_{k-1}+1), \quad (3.5)
 \end{aligned}$$

where $g(\tilde{x}) = r_j$, by the monotone likelihood ratio property of the beta distribution ($I_x(a, b)$ has a monotone likelihood ratio in $-a$ for fixed b , which implies that $I_x(a, b)$ is a decreasing function of a). A similar argument applies to the other indexes. Thus, if (3.4) is true for $n_1 = n_2 = \dots = n_k$, by (3.5) it follows for $n_1 \leq n_2 \leq \dots \leq n_k$.

So, assuming $\tilde{n} = (n_1, n_1, \dots, n_1)$, we seek to maximize

$$P_p\left\{\sum_{i=1}^k \sum_{j=1}^{n_1} Y_{ij} \geq \sum_{i=1}^k (n_i - x_i) = \sum_{i=1}^k y_i\right\}, \quad (3.6)$$

where Y_{ij} are independent Bernoulli random variables with

parameter p_i and $\prod_{i=1}^k p_i^{y_i} = a$. If $\prod_{i=1}^k p_i^{y_i} = a$, then $\prod_{i=1}^k p_i$

ranges from a^{1/γ_j} to a^{1/γ_1} , $\gamma_j = \min_{1 \leq i \leq k} \gamma_i$. This is seen as follows:

$$\begin{aligned} \prod_{i=1}^k p_i &= \left(\prod_{i=1}^k p_i^{\gamma_i} \right)^{1/\gamma_1} \prod_{i=2}^k p_i^{1-\gamma_i/\gamma_1} \\ &= a^{1/\gamma_1} \prod_{i=2}^k p_i^{(\gamma_1 - \gamma_i)/\gamma_1} \leq a^{1/\gamma_1} \end{aligned}$$

and

$$\begin{aligned} \prod_{i=1}^k p_i &= \left(\prod_{i=1}^k p_i^{\gamma_i} \right)^{1/\gamma_j} \prod_{i \neq j}^k p_i^{1-\gamma_i/\gamma_j} \\ &= a^{1/\gamma_j} \prod_{i \neq j}^k p_i^{(\gamma_j - \gamma_i)/\gamma_j} > a^{1/\gamma_j} \end{aligned}$$

and the choices $p_1 = a^{1/\gamma_1}$, $p_2 = \dots = p_k = 1$, and $p_j = a^{1/\gamma_j}$, $p_i = 1$, $i \neq j$, attain these values. From the results of Pledger

and Proschan (1971), for each $b = \prod_{i=1}^k p_i$, $a^{1/\gamma_j} < b < a^{1/\gamma_1}$,

(3.6) is maximized by $p_1 = b$, $p_i = 1$, $2 \leq i \leq k$. Further, the

maximum over b , $a^{1/\gamma_j} < b < a^{1/\gamma_1}$, of the maxima for each b is

given by $p_1 = a^{1/\gamma_1}$, $p_i = 1$, $2 \leq i \leq k$, by the monotone likelihood ratio property of the binomial distribution, and

$p_1 = a^{1/\gamma_1}$, $p_i = 1$, $2 \leq i \leq k$, satisfies $\prod_{i=1}^k p_i^{\gamma_i} = a$. This completes the proof.

If $\gamma_1 = \gamma_2 = \dots = \gamma_k = 1$, some guidelines for the verification of (3.3) are given in Harris and Soms (1983). In the present case (3.3) must be verified by trial and error by showing

that $\min_{\sum_{i=1}^k x_i = x_1} \prod_{i=1}^k (n_i - x_i)^{\gamma_i} = (n_1 - x_1)^{\gamma_1} \prod_{i=2}^k n_i^{\gamma_i}$ and that

$$\max_{\sum_{i=1}^k x_i = x_1 + 1} \prod_{i=1}^k (n_i - x_i)^{\gamma_i} < (n_1 - x_1)^{\gamma_1} \prod_{i=2}^k n_i^{\gamma_i}.$$

Example 3.1. Let $k = 3$, $\tilde{n} = (5, 5, 5)$, $\tilde{\gamma} = (3, 3, 2)$, $\alpha = .10$ and

$$\tilde{x} = (1, 0, 0). \text{ Then } \min_{\sum_{i=1}^3 x_i = 1} \prod_{i=1}^3 (n_i - x_i)^{\gamma_i} = 200000 \text{ and}$$

$$\max_{\sum_{i=1}^3 x_i = 2} \prod_{i=1}^3 (n_i - x_i)^{\gamma_i} = 140625 \text{ and } \tilde{x} \text{ is a key test result}$$

and (3.3) is satisfied and hence

$$a_{g(\tilde{x})} = .4161^3 = .0720 ,$$

where $.10 = I_{.4161}(4, 2)$. Further, it can also be verified that $\tilde{x} = (2, 0, 0)$ is a key test result for which (3.3) is satisfied, but that for $\tilde{x} = (3, 0, 0)$, (3.3) is violated.

Note that Theorem 3.2 asserts that $a_{g(\tilde{x})} = b_{\alpha}^{\gamma_1}$ for $0 < \alpha < 1$. It is thus possible that (3.3) is not true but the conclusion still holds for α of practical importance. This is taken up in Section 4.

4. THE LINDSTROM-MADDEN METHOD FOR SERIES SYSTEMS WITH REPEATED COMPONENTS

When $\gamma_1 = \gamma_2 = \dots = \gamma_r = 1$, the Lindstrom-Madden method (henceforth abbreviated L-M) is an approximation $b_{g(\tilde{x})}$ to $a_{g(\tilde{x})}$ of the form

$$b_{g(\tilde{x})} = \min_{1 \leq i \leq k} b_{\alpha}(n_i) , \quad (4.1)$$

where

$$\alpha = I_{b_{\alpha}(n_i)}(n_i - t_{0i}, t_{0i} + 1), \quad (4.2)$$

with

$$t_{0i} = n_i \left(1 - \prod_{j=1}^k (n_j - x_j)/n_i\right), \quad (4.3)$$

i.e., t_{0i} is the maximum of the recursive indexes t_i defined by (2.4). For the usual levels of α , $b_{g(\tilde{x})} = b_{\alpha}(n_1)$. Further, numerical evidence indicates (Harris and Soms (1983)) that for α levels of practical significance

$$b_{g(\tilde{x})} \leq a_{g(\tilde{x})}. \quad (4.4)$$

(4.4) was incorrectly claimed to be true for $0 < \alpha < 1$ in Sudakov (1974) and this is discussed at length in Harris and Soms (1983). However, (4.4) is known to hold for special cases (Winterbottom (1974) and Harris and Soms (1983)).

Motivated by the above, we now give an L-M approximation $b_{g(\tilde{x})}$ to $a_{g(\tilde{x})}$ for arbitrary γ_i by

$$b_{g(\tilde{x})} = \min_{1 \leq i \leq k} b_{\alpha}(n_i)^{\gamma_i}, \quad (4.5)$$

where

$$\alpha = I_{b_{\alpha}(n_i)}(n_i - t_{0i}, t_{0i} + 1), \quad (4.6)$$

with

$$t_{0i} = n_i - \left(\prod_{j=1}^k (n_j - x_j)^{\gamma_j} \right)^{1/\gamma_i}, \quad (4.7)$$

i.e., t_{0i} is the maximum of the recursive indexes t_i defined by (2.4). However, in this case it is not clear which index i gives the minimum, except that the likely candidate is the one for

which γ_j , $1 \leq j \leq k$, is a maximum. We might expect, by analogy, that for α levels of practical interest

$$b_{g(\tilde{x})} \leq a_{g(\tilde{x})} . \quad (4.8)$$

5. NUMERICAL EXAMPLES

For $k = 2$ and selected $\tilde{n}, \tilde{\gamma}, \tilde{x}$, $\alpha = .05$ and $.10$, Table I gives $b_{g(\tilde{x})}$, $a_{g(\tilde{x})}$ and the best upper bound, $u_{g(\tilde{x})}$,

$$u_{g(\tilde{x})} = \min_{1 \leq i \leq k} u_{\alpha}(n_i)^{\gamma_i}, \quad (5.1)$$

where

$$\alpha = I_{u_{\alpha}(n_i)}(n_i - [t_{0i}], [t_{0i}] + 1) \quad (5.2)$$

and t_{0i} are defined as in (4.6).

TABLE I.

L-M Approximations and $a_{g(\tilde{x})}$

(n_1, n_2)	(γ_1, γ_2)	(x_1, x_2)	α	$b_{g(\tilde{x})}$	$a_{g(\tilde{x})}$	$u_{g(\tilde{x})}$
(10, 10)	(1, 2)	(0, 1)	.05	.3670	.3670	.3670
(10, 10)	(1, 2)	(0, 1)	.10	.4398	.4398	.4398
(10, 10)	(1, 2)	(1, 1)	.05	.3045	.3514	.3670
(10, 10)	(1, 2)	(1, 1)	.10	.3715	.4227	.4398
(10, 10)	(1, 2)	(2, 1)	.05	.2484	.3151	.3670
(10, 10)	(1, 2)	(2, 1)	.10	.3088	.3825	.4398
(10, 15)	(2, 3)	(0, 1)	.05	.3695	.3719	.3742
(10, 15)	(2, 3)	(0, 1)	.10	.4425	.4446	.4467
(10, 15)	(2, 3)	(1, 1)	.05	.2554	.3042	.3670
(10, 15)	(2, 3)	(1, 1)	.10	.3167	.3705	.4398
(10, 15)	(2, 3)	(2, 1)	.05	.1712	.1981	.2431
(10, 15)	(2, 3)	(2, 0)	.10	.2203	.2513	.3029

Note that for all the cases in Table I, $b_{g(x)}$ is a lower bound for $a_{g(x)}$. The computations were done by a short FORTRAN program, a listing of which can be obtained from the author.

6. CONCLUDING REMARKS

In this paper we have extended the L-M method to series systems with repeated components. More work is needed to ascertain the region of validity of (4.8).

7. ACKNOWLEDGEMENTS

This work was sponsored in part by the United States Army under Contract No. DAAG29-80-C-0041, the Office of Naval Research under Contract No. N00014-79-C-0321, and the University of Wisconsin-Milwaukee.

BIBLIOGRAPHY

- Buehler, R. J., (1957). Confidence Limits for the Product of Two Binomial Parameters. Journal of the American Statistical Association, 52, 482-93.
- Harris, B. and Soms, A. P., (1983). The Theory of Optimal Confidence Limits for Systems Reliability with Counterexamples for Results on Optimal Confidence Limits for Series Systems. Technical Report #708, Department of Statistics, University of Wisconsin-Madison.
- Pledger, G. and Proschan, F., (1971). Comparison of Order Statistics and of Spacings from Heterogeneous Distributions. Optimizing Methods in Statistics, New York: Academic Press, 89-113.
- Sudakov, R. S., (1974). On the Question of Interval Estimation of the Index of Reliability of a Sequential System. Engineering Cybernetics, 12, 55-63.
- Winterbottom, A., (1974). Lower Limits for Series System Reliability from Binomial Data. Journal of the American Statistical Association, 69, 782-8.

How to Display Data Badly

HOWARD WAINER*

Methods for displaying data badly have been developing for many years, and a wide variety of interesting and inventive schemes have emerged. Presented here is a synthesis yielding the 12 most powerful techniques that seem to underlie many of the realizations found in practice. These 12 (the dirty dozen) are identified and illustrated.

KEY WORDS: Graphics; Data display; Data density; Data-ink ratio.

1. INTRODUCTION

The display of data is a topic of substantial contemporary interest and one that has occupied the thoughts of many scholars for almost 200 years. During this time there have been a number of attempts to codify standards of good practice (e.g., ASME Standards 1915; Cox 1978; Ehrenberg 1977) as well as a number of books that have illustrated them (i.e., Bertin 1973, 1977, 1981; Schmid 1954; Schmid and Schmid 1979; Tufte 1983). The last decade or so has seen a tremendous increase in the development of new display techniques and tools that have been reviewed recently (Macdonald-Ross 1977; Fienberg 1979; Cox 1978; Wainer and Thissen 1981). We wish to concentrate on methods of data display that leave the viewers as uninformed as they were before seeing the display or, worse, those that induce confusion. Although such techniques are broadly practiced, to my knowledge they have not as yet been gathered into a single source or carefully

*Howard Wainer is Senior Research Scientist, Educational Testing Service, Princeton, NJ 08541. This is the text of an invited address to the American Statistical Association. It was supported in part by the Program Statistics Research Project of the Educational Testing Service. The author would like to express his gratitude to the numerous friends and colleagues who read or heard this article and offered valuable suggestions for its improvement. Especially helpful were David Andrews, Paul Holland, Bruce Kaplan, James O. Ramsay, Edward Tufte, the participants in the Stanford Workshop on Advanced Graphical Presentation, two anonymous referees, the long-suffering associate editor, and Gary Koch.

categorized. This article is the beginning of such a compendium.

The aim of good data graphics is to display data accurately and clearly. Let us use this definition as a starting point for categorizing methods of bad data display. The definition has three parts. These are (a) showing data, (b) showing data accurately, and (c) showing data clearly. Thus, if we wish to display data badly, we have three avenues to follow. Let us examine them in sequence, parse them into some of their component parts, and see if we can identify means for measuring the success of each strategy.

2. SHOWING DATA

Obviously, if the aim of a good display is to convey information, the less information carried in the display,

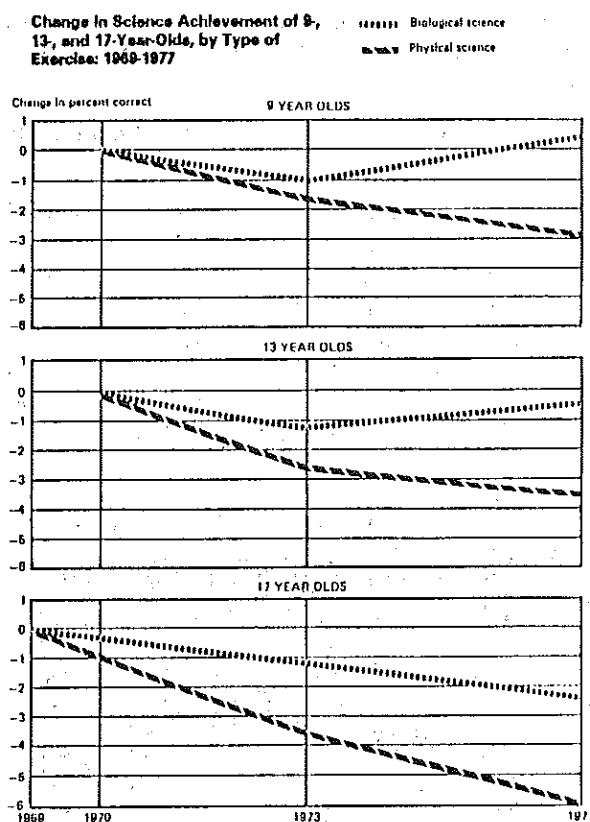


Figure 1. An example of a low density graph (from SI3 [ddi = .3]).

© The American Statistician, May 1984, Vol. 38, No. 2

This article appeared in the American Statistician Vol. 38, No. 2, pp 137-147. Permission of the author and the editor of that journal to reproduce it here is appreciated.

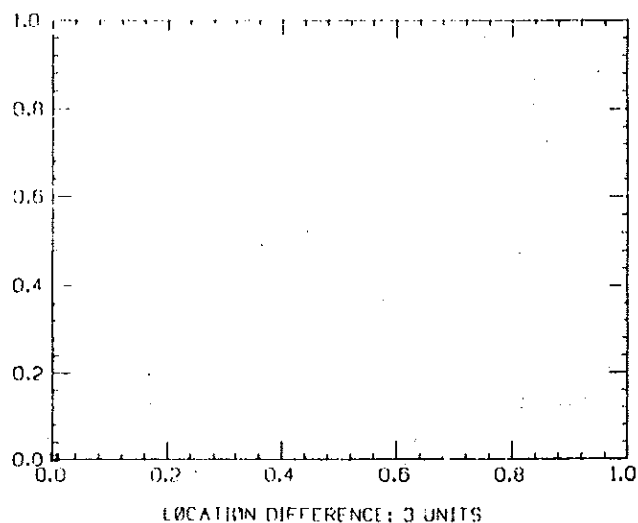


Figure 2. A low density graph (from Friedman and Rafsky 1981 (ddi = .5)).

the worse it is. Tufte (1983) has devised a scheme for measuring the amount of information in displays, called the data density index (ddi), which is "the number of numbers plotted per square inch." This easily calculated index is often surprisingly informative. In popular and technical media we have found a range from .1 to 362. This provides us with the first rule of bad data display.

Rule 1—Show as Few Data as Possible (Minimize the Data Density)

What does a data graphic with a ddi of .3 look like? Shown in Figure 1 is a graphic from the book *Social Indicators III* (SI3), originally done in four colors (original size 7" by 9") that contains 18 numbers ($18/63 = .3$). The median data graph in SI3 has a data density of .6 numbers/in²; this one is not an unusual choice. Shown in Figure 2 is a plot from the article by Friedman and Rafsky (1981) with a ddi of .5 (it shows 4 numbers in 8

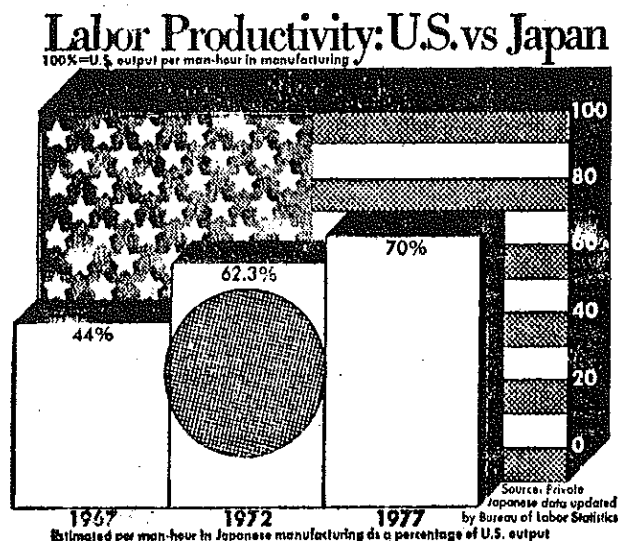


Figure 3. A low density graph (© 1978, The Washington Post) with chart-junk to fill in the space (ddi = .2).

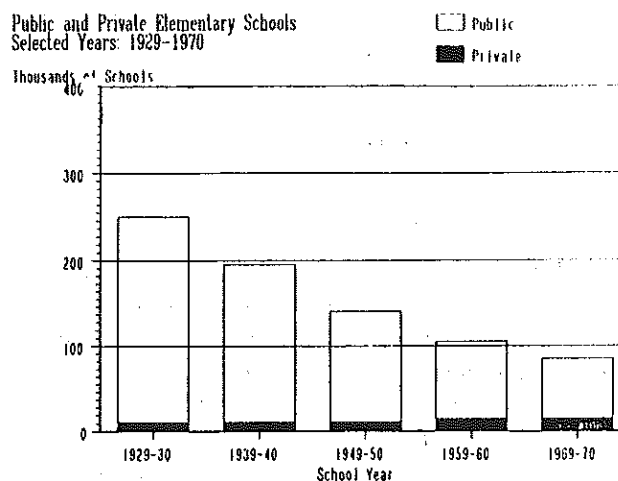


Figure 4. Hiding the data in the scale (from SI3).

in²). This is unusual for JASA, where the median data graph has a ddi of 27. In defense of the producers of this plot, the point of the graph is to show that a method of analysis suggested by a critic of their paper was not fruitful. I suspect that prose would have worked pretty well also.

Although arguments can be made that high data density does not imply that a graphic will be good, nor one with low density bad, it does reflect on the efficiency of the transmission of information. Obviously, if we hold clarity and accuracy constant, more information is bet-

THE NUMBER OF PRIVATE ELEMENTARY SCHOOLS FROM 1930-1970

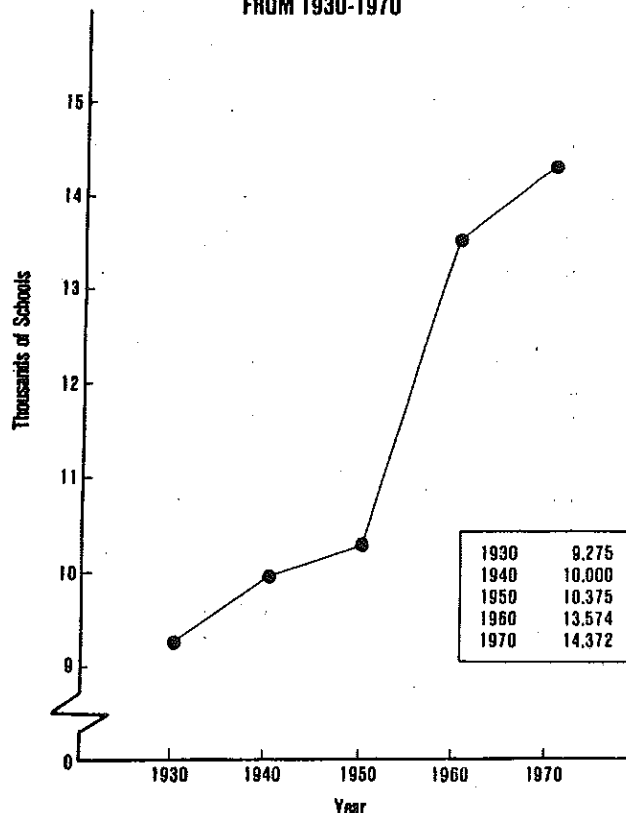


Figure 5. Expanding the scale and showing the data in Figure 4 (from SI3).

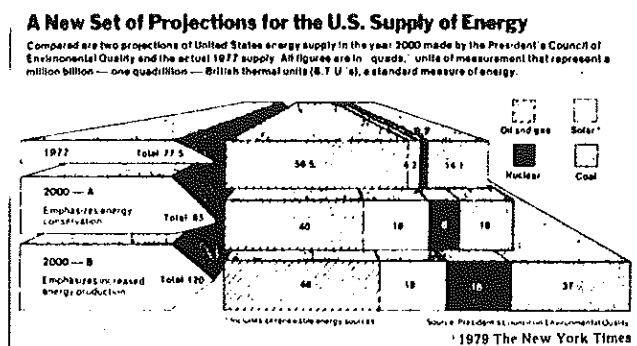


Figure 6. Ignoring the visual metaphor (© 1978, The New York Times).

ter than less. One of the great assets of graphical techniques is that they can convey large amounts of information in a small space.

We note that when a graph contains little or no information the plot can look quite empty (Figure 2) and thus raise suspicions in the viewer that there is nothing to be communicated. A way to avoid these suspicions is to fill up the plot with nondata figurations—what Tufte has termed “chartjunk.” Figure 3 shows a plot of the labor productivity of Japan relative to that of the United States. It contains one number for each of three years. Obviously, a graph of such sparse information would have a lot of blank space, so filling the space hides the paucity of information from the reader.

A convenient measure of the extent to which this practice is in use is Tufte’s “data-ink ratio.” This measure is the ratio of the amount of ink used in graphing the data to the total amount of ink in the graph. The closer to zero this ratio gets, the worse the graph. The notion of the data-ink ratio brings us to the second principle of bad data display.

Rule 2—Hide What Data You Do Show (Minimize the Data-Ink Ratio)

One can hide data in a variety of ways. One method that occurs with some regularity is hiding the data in the grid. The grid is useful for plotting the points, but only rarely afterwards. Thus to display data badly, use a fine grid and plot the points dimly (see Tufte 1983, pp. 94–95 for one repeated version of this).

A second way to hide the data is in the scale. This corresponds to blowing up the scale (i.e., looking at the data from far away) so that any variation in the data is obscured by the magnitude of the scale. One can justify this practice by appealing to “honesty requires that we start the scale at zero,” or other sorts of sophistry.

In Figure 4 is a plot that (from S13) effectively hides the growth of private schools in the scale. A redrawing of the number of private schools on a different scale conveys the growth that took place during the mid-1950’s (Figure 5). The relationship between this rise and *Brown vs. Topeka School Board* becomes an immediate question.

To conclude this section, we have seen that we can display data badly either by not including them (Rule 1)

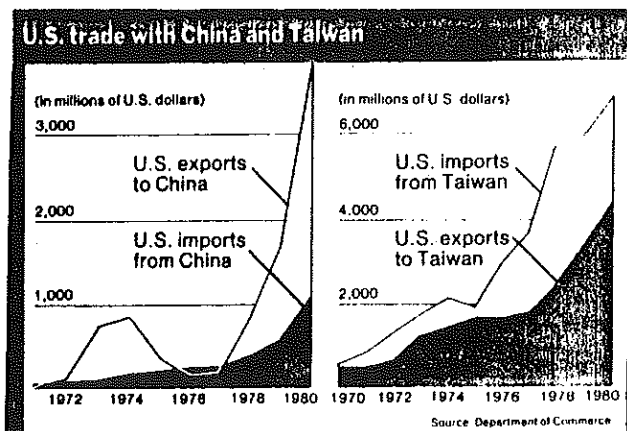


Figure 7. Reversing the metaphor in mid-graph while changing scales on both axes (© June 14, 1981, The New York Times).

or by hiding them (Rule 2). We can measure the extent to which we are successful in excluding the data through the data density; we can sometimes convince viewers that we have included the data through the incorporation of chartjunk. Hiding the data can be done either by using an overabundance of chartjunk or by cleverly choosing the scale so that the data disappear. A measure of the success we have achieved in hiding the data is through the data-ink ratio.

3. SHOWING DATA ACCURATELY

The essence of a graphic display is that a set of numbers having both magnitudes and an order are represented by an appropriate visual metaphor—the magnitude and order of the metaphorical representation match the numbers. We can display data badly by ignoring or distorting this concept.

Rule 3—Ignore the Visual Metaphor Altogether

If the data are ordered and if the visual metaphor has a natural order, a bad display will surely emerge if you shuffle the relationship. In Figure 6 note that the bar labeled 14.1 is longer than the bar labeled 18. Another method is to change the meaning of the metaphor in the middle of the plot. In Figure 7 the dark shading represents imports on one side and exports on the other. This is but one of the problems of this graph; more serious still is the change of scale. There is also a difference in the time scale, but that is minor. A common theme in Playfair’s (1786) work was the difference between imports and exports. In Figure 8, a 200-year-old graph tells the story clearly. Two such plots would have illustrated the story surrounding this graph quite clearly.

Rule 4—Only Order Matters

One frequent trick is to use length as the visual metaphor when area is what is perceived. This was used quite effectively by *The Washington Post* in Figure 9. Note that this graph also has a low data density (.1), and its data-ink ratio is close to zero. We can also calculate Tufte’s (1983) measure of perceptual distortion (PD) for this graph. The PD in this instance is the perceived

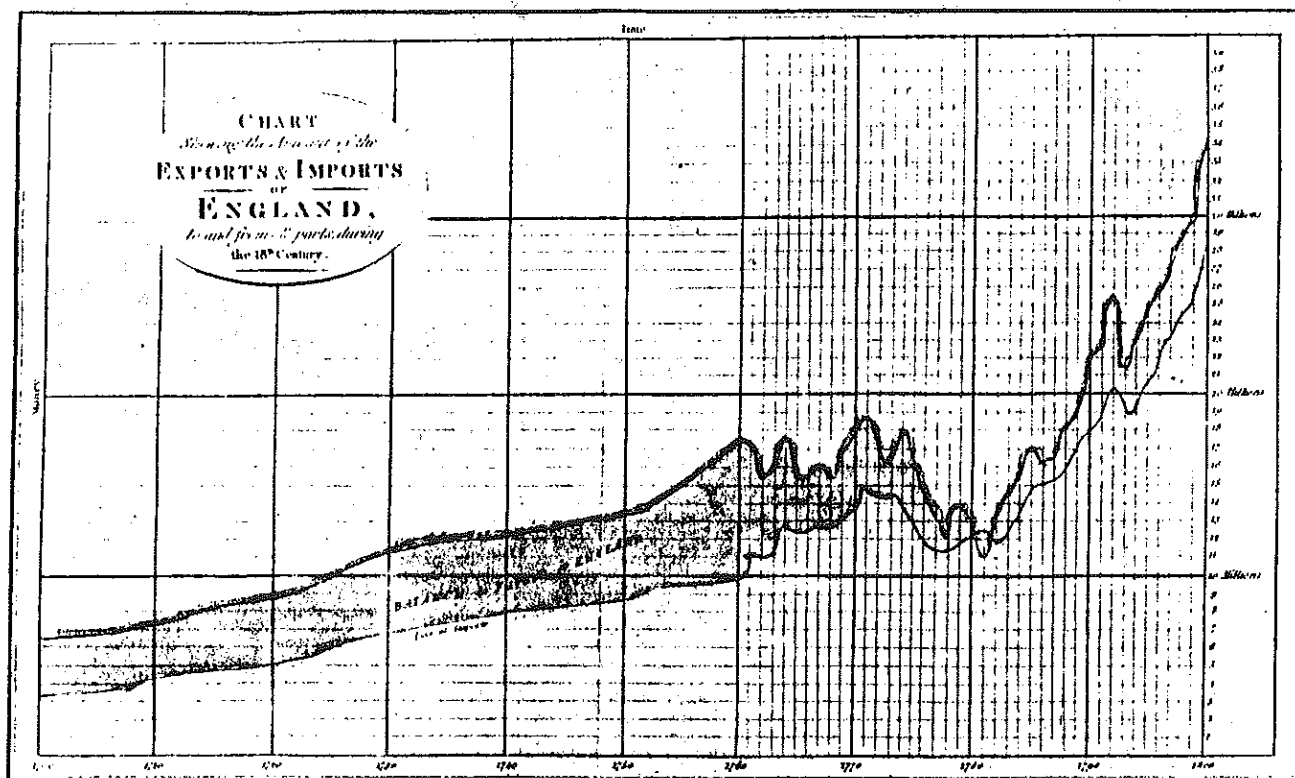


Figure 8. A plot on the same topic done well two centuries earlier (from Playfair 1786).

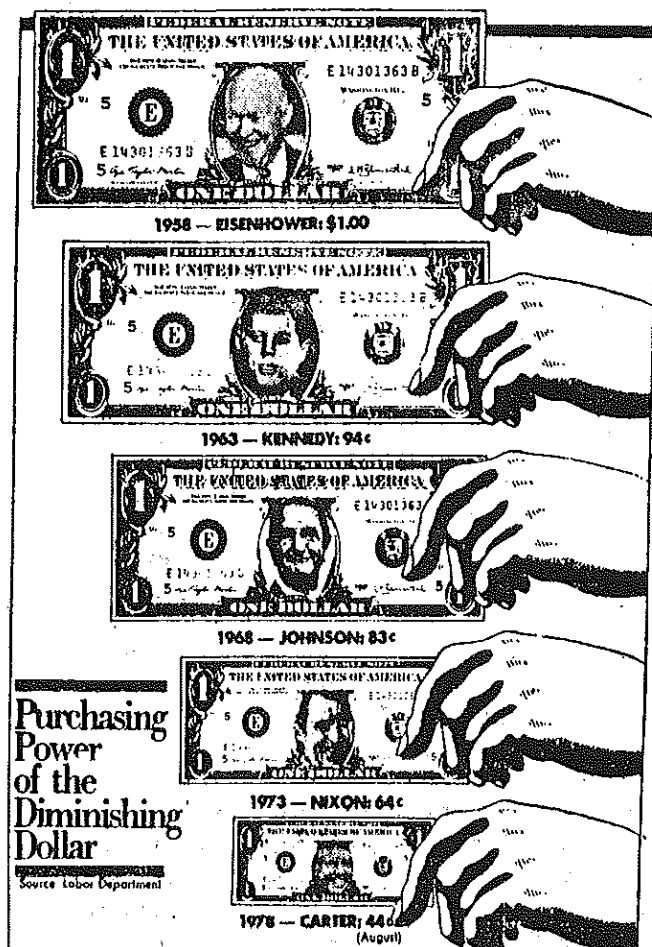


Figure 9. An example of how to goose up the effect by squaring the eyeball (© 1978, The Washington Post).

change in the value of the dollar from Eisenhower to Carter divided by the actual change. I read and measure thus:

$$\frac{\text{Actual}}{1.00 - .44} = 1.27 \quad \frac{\text{Measured}}{22.00 - 2.06} = 9.68$$

$$PD = 9.68/1.27 = 7.62$$

This distortion of over 700% is substantial but by no means a record.

A less distorted view of these data is provided in Figure 10. In addition, the spacing suggested by the

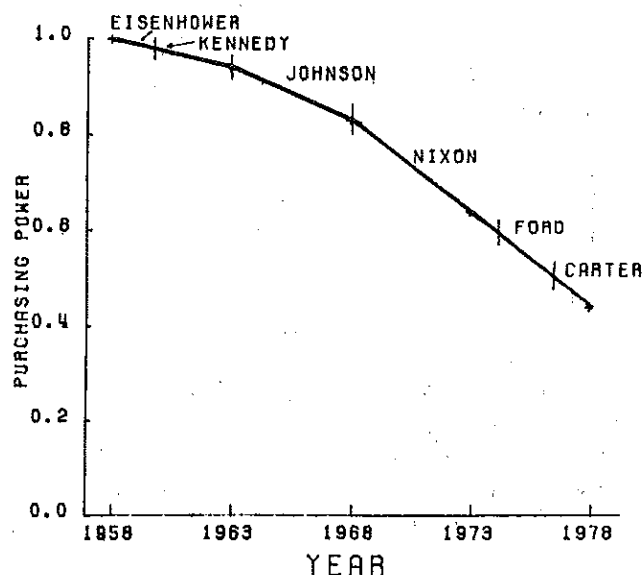


Figure 10. The data in Figure 9 as an unadorned line chart (from Walner, 1980).

presidential faces is made explicit on the time scale.

Rule 5—Graph Data Out of Context

Often we can modify the perception of the graph (particularly for time series data) by choosing carefully the interval displayed. A precipitous drop can disappear if we choose a starting date just after the drop. Similarly, we can turn slight meanders into sharp changes by focusing on a single meander and expanding the scale. Often the choice of scale is arbitrary but can have profound effects on the perception of the display. Figure 11 shows a famous example in which President Reagan gives an out-of-context view of the effects of his tax cut. The *Times*' alternative provides the context for a deeper understanding. Simultaneously omitting the context as well as any quantitative scale is the key to the practice of Ordinal Graphics (see also Rule 4). Automatic rules do not always work, and wisdom is always required.

In Section 3 we discussed three rules for the accurate display of data. One can compromise accuracy by ignoring visual metaphors (Rule 3), by only paying attention to the order of the numbers and not their magnitude (Rule 4), or by showing data out of context (Rule 5). We advocated the use of Tufte's measure of perceptual distortion as a way of measuring the extent to which the accuracy of the data has been compromised by the display. One can think of modifications that would allow it to be applied in other situations, but we leave such expansion to other accounts.

4. SHOWING DATA CLEARLY

In this section we discuss methods for badly displaying data that do not seem as serious as those de-

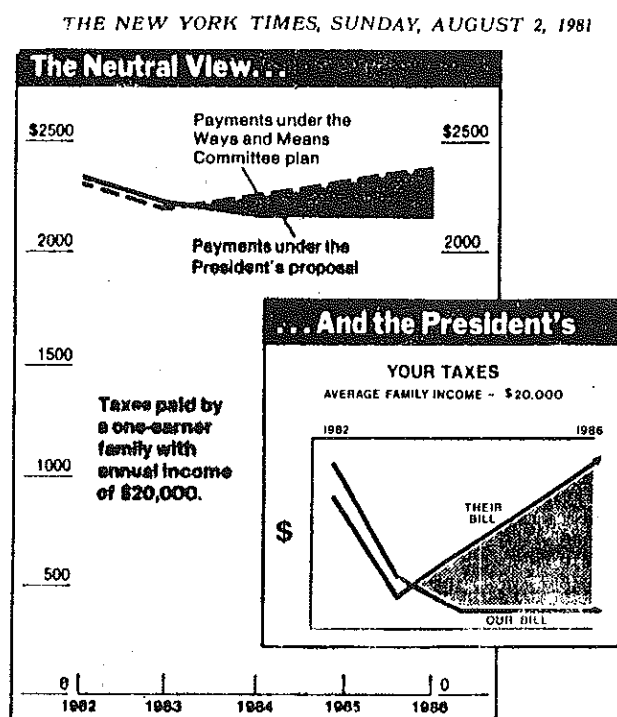


Figure 11. The White House showing neither scale nor context (© 1981, The New York Times, reprinted with permission).

scribed previously; that is, the data are displayed, and they might even be accurate in their portrayal. Yet subtle (and not so subtle) techniques can be used to effectively obscure the most meaningful or interesting aspects of the data. It is more difficult to provide objective measures of presentational clarity, but we rely on the reader to judge from the examples presented.

Rule 6—Change Scales in Mid-Axis

This is a powerful technique that can make large differences look small and make exponential changes look linear.

In Figure 12 is a graph that supports the associated story about the skyrocketing circulation of *The New York Post* compared to the plummeting *Daily News* circulation. The reason given is that New Yorkers "trust" the *Post*. It takes a careful look to note the 700,000 jump that the scale makes between the two lines.

In Figure 13 is a plot of physicians' incomes over time. It appears to be linear, with a slight tapering off in recent years. A careful look at the scale shows that it starts out plotting every eight years and ends up plotting yearly. A more regular scale (in Figure 14) tells quite a different story.

The soaraway Post — the daily paper New Yorkers trust

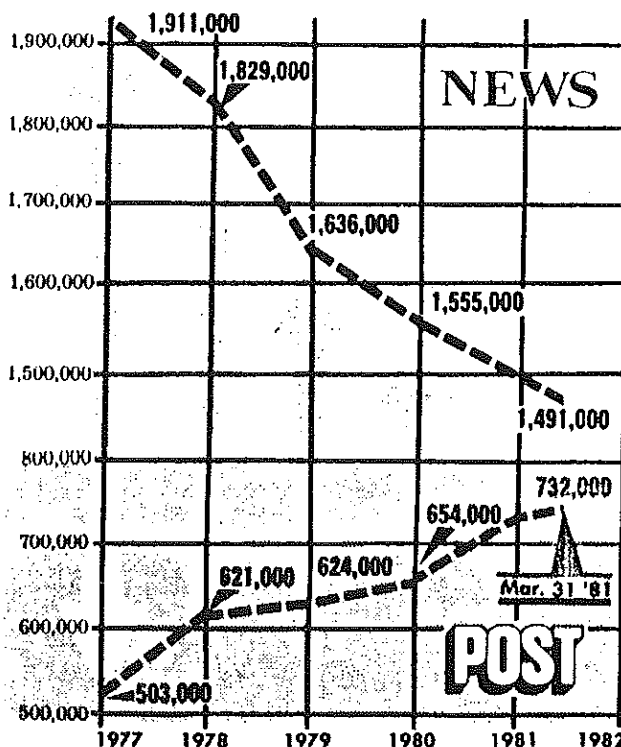


Figure 12. Changing scale in mid-axis to make large differences small (© 1981, New York Post).

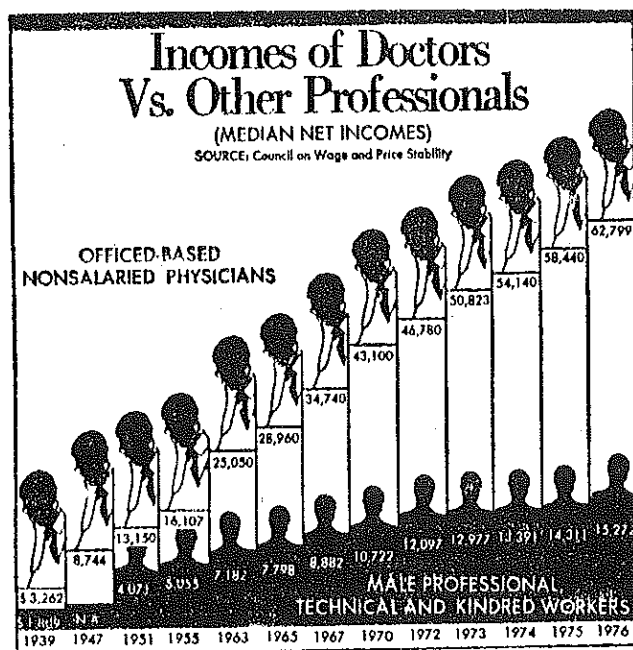


Figure 13. Changing scale in mid-axis to make exponential growth linear (© The Washington Post).

Rule 7—Emphasize the Trivial (Ignore the Important)

Sometimes the data that are to be displayed have one important aspect and others that are trivial. The graph can be made worse by emphasizing the trivial part. In Figure 15 we have a page from *SL3* that compares the income levels of men and women by educational levels. It reveals the not surprising result that better educated individuals are paid better than more poorly educated ones and that changes across time expressed in constant dollars are reasonably constant. The comparison of greatest interest and current concern, comparing salaries between sexes within education level, must be made clumsily by vertically transposing from one graph to another. It seems clear that Rule 7 must have been operating here, for it would have been easy to place the graphs side by side and allow the comparison of interest to be made more directly. Looking at the problem from a strictly data-analytic point of view, we note that there are two large main effects (education and sex) and a small time effect. This would have implied a plot that

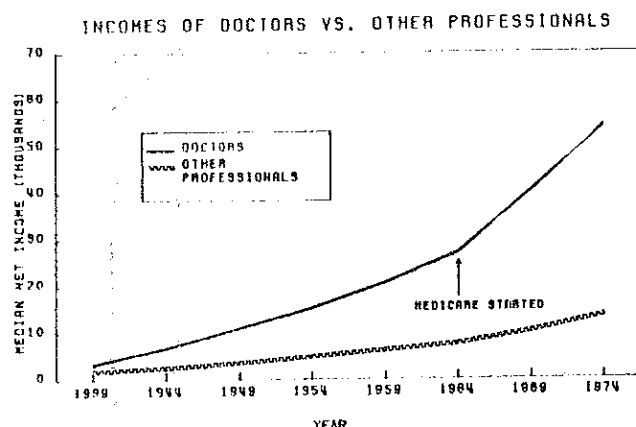


Figure 14. Data from Figure 13 redone with linear scale (from Wainer 1980).

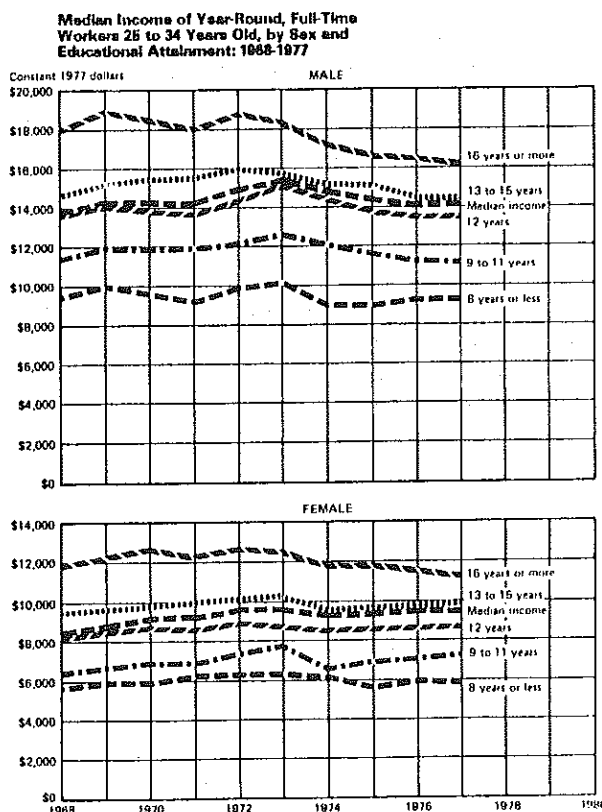


Figure 15. Emphasizing the trivial: Hiding the main effect of sex differences in income through the vertical placement of plots (from *SL3*).

showed the large effects clearly and placed the smallish time trend into the background (Figure 16).

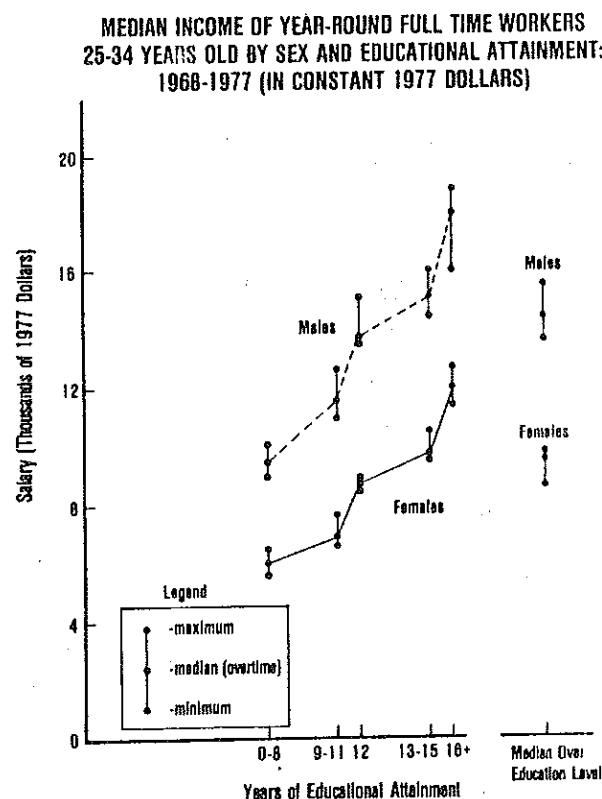


Figure 16. Figure 15 redone with the large main effects emphasized and the small one (time trends) suppressed.

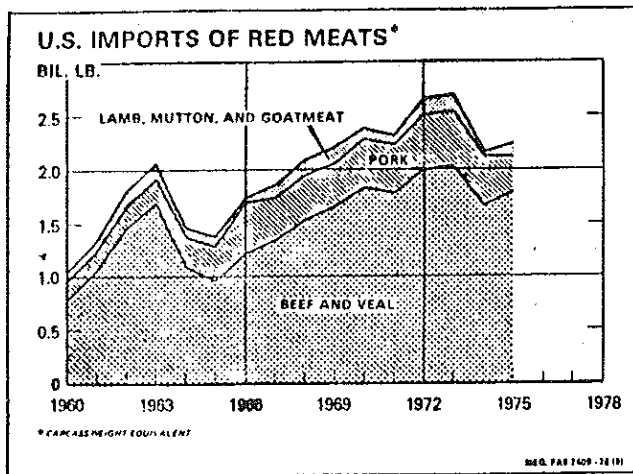
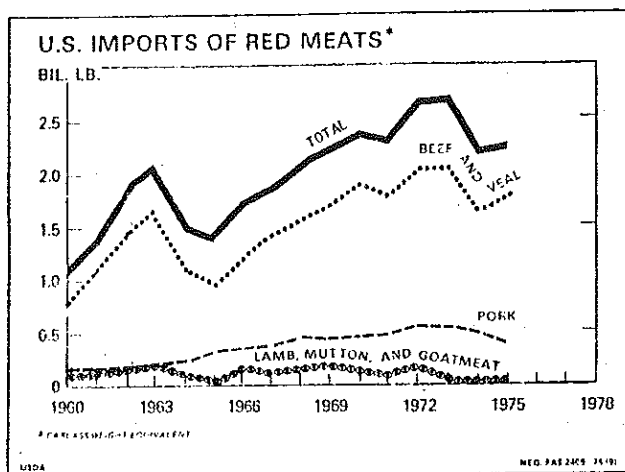


Figure 17. Jiggling the baseline makes comparisons more difficult (from Handbook of Agricultural Charts).

Rule 8—Jiggle the Baseline

Making comparisons is always aided when the quantities being compared start from a common base. Thus we can always make the graph worse by starting from different bases. Such schemes as the hanging or suspended rootogram and the residual plot are meant to facilitate comparisons. In Figure 17 is a plot of U.S. imports of red meat taken from the *Handbook of Agricultural Charts* published by the U.S. Department of Agriculture. Shading beneath each line is a convention that indicates summation, telling us that the amount of each kind of meat is added to the amounts below it. Because of the dominance of and the fluctuations in importation of beef and veal, it is hard to see what the changes are in the other kinds of meat—Is the importation of pork increasing? Decreasing? Staying constant? The only purpose for stacking is to indicate graphically the total summation. This is easily done through the addition of another line for TOTAL. Note that a TOTAL will always be clear and will never intersect the other lines on the plot. A version of these data is shown



Source: Handbook of Agricultural Charts, U.S. Department of Agriculture, 1976, p. 93.
Chart Source: Original

Figure 18. An alternative version of Figure 17 with a straight line used as the basis of comparison.

Life Expectancy at Birth, by Sex, Selected Countries, Most Recent Available Year: 1970-1975

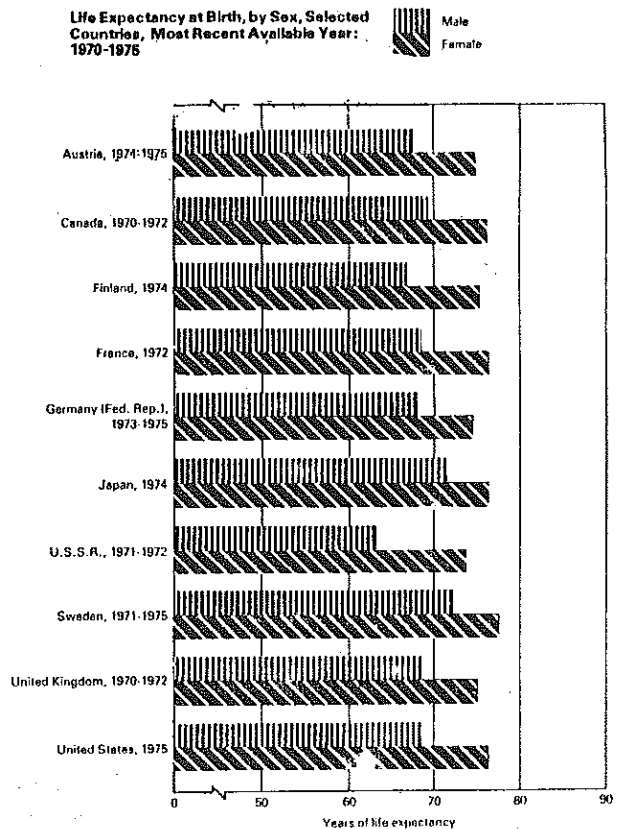


Figure 19. Austria First! Obscuring the data structure by alphabetizing the plot (from SI3).

in Figure 18 with the separate amounts of each meat, as well as a summation line, shown clearly. Note how easily one can see the structure of import of each kind of meat now that the standard of comparison is a straight line (the time axis) and no longer the import amount of those meats with greater volume.

Rule 9—Austria First!

Ordering graphs and tables alphabetically can obscure structure in the data that would have been obvious had the display been ordered by some aspect of the data. One can defend oneself against criticisms by pointing out that alphabetizing "aids in finding entries of interest." Of course, with lists of modest length such aids are unnecessary; with longer lists the indexing schemes common in 19th century statistical atlases provide easy lookup capability.

Figure 19 is another graph from SI3 showing life expectancies, divided by sex, in 10 industrialized nations. The order of presentation is alphabetical (with the USSR positioned as Russia). The message we get is that there is little variation and that women live longer than men. Redone as a stem-and-leaf diagram (Figure 20 is simply a reordering of the data with spacing proportional to the numerical differences), the magnitude of the sex difference leaps out at us. We also note that the USSR is an outlier for men.

Rule 10—Label (a) Illegibly, (b) Incompletely, (c) Incorrectly, and (d) Ambiguously

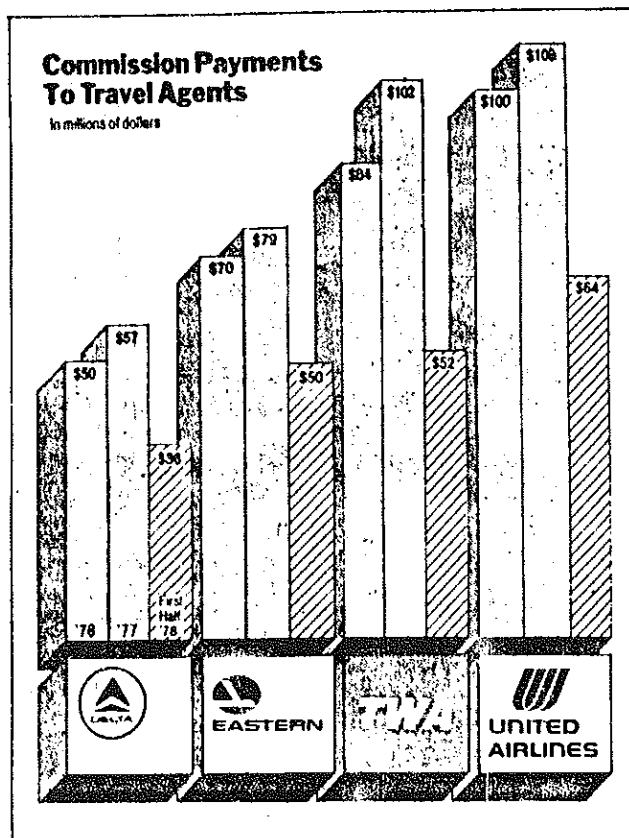
There are many instances of labels that either do not

LIFE EXPECTANCY AT BIRTH, BY SEX,
MOST RECENT AVAILABLE YEAR

WOMEN	YEARS	MEN
SWEDEN	78	
	77	
FRANCE, US, JAPAN, CANADA	76	
FINLAND, AUSTRIA, UK	75	
USSR, GERMANY	74	
	73	
	72	SWEDEN
	71	JAPAN
	70	
	69	CANADA, UK, US, FRANCE
	68	GERMANY, AUSTRIA
	67	FINLAND
	66	
	65	
	64	
	63	USSR
	62	

Figure 20. Ordering and spacing the data from Figure 19 as a stem-and-leaf diagram provides insights previously difficult to extract (from SI3).

tell the whole story, tell the wrong story, tell two or more stories, or are so small that one cannot figure out what story they are telling. One of my favorite examples of small labels is from *The New York Times* (August



Complex web of discount fares and airlines' telephone delays are raising travel agents' overhead, offsetting revenue gains from higher volume.

Figure 21. Mixing a changed metaphor with a tiny label reverses the meaning of the data (© 1978, The New York Times).

Commission Payments to Travel Agents

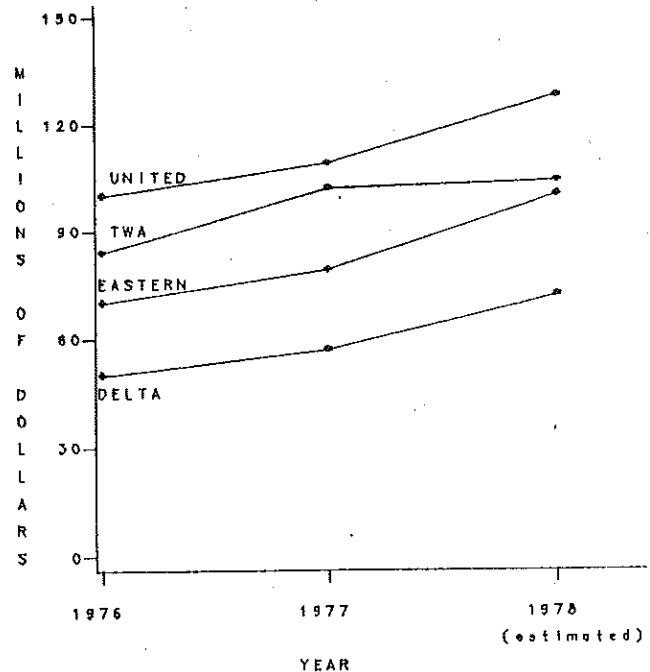


Figure 22. Figure 21 redrawn with 1978 data placed on a comparable basis (from Wainer 1980).

1978), in which the article complains that fare cuts lower commission payments to travel agents. The graph (Figure 21) supports this view until one notices the tiny label indicating that the small bar showing the decline is for just the first half of 1978. This omits such heavy travel periods as Labor Day, Thanksgiving, Christmas, and so on, so that merely doubling the first-half data is probably not enough. Nevertheless, when this bar is doubled (Figure 22), we see that the agents are doing very well indeed compared to earlier years.

Rule 11—More Is Murkier: (a) More Decimal Places and (b) More Dimensions

We often see tables in which the number of decimal places presented is far beyond the number that can be perceived by a reader. They are also commonly presented to show more accuracy than is justified. A display can be made clearer by presenting less. In Table 1 is a section of a table from Dhariyal and Dudewicz's (1981) JASA paper. The table entries are presented to five decimal places! In Table 2 is a heavily rounded version that shows what the authors intended clearly. It also shows that the various columns might have a substantial redundancy in them (the maximum expected gain with $b/c = 10$ is about $1/10$ th that of $b/c = 100$ and $1/100$ th that of $b/c = 1,000$). If they do, the entire table could have been reduced substantially.

Just as increasing the number of decimal places can make a table harder to understand, so can increasing the number of dimensions make a graph more con-

Table 1. Optimal Selection From a Finite Sequence With Sampling Cost

N	b/c = 10.0		100.0		1,000.0	
	r*	(G _N (r*) - a)/c	r*	(G _N (r*) - a)/c	r*	(G _N (r*) - a)/c
3	2	.20000	2	2.22500	2	22.47499
4	2	.26333	2	2.88833	2	29.13832
5	2	.32333	3	3.54167	3	35.79166
6	3	.38267	3	4.23767	3	42.78764
7	3	.44600	3	4.90100	3	49.45097
8	3	.50743	4	5.57650	4	56.33005
9	3	.56743	4	6.26025	4	63.20129
10	4	.62948	4	6.92358	4	69.86462

NOTE: $g(X_s | r-1) = bR(X_s | r-1) + a$, if $S = s$, and $g(X_s | r-1) = 0$, otherwise.
Source: Dhariyal and Dudewicz (1981).

fusing. We have already seen how extra dimensions can cause ambiguity (Is it length or area or volume?). In addition, human perception of areas is inconsistent. Just what is confusing and what is not is sometimes only a conjecture, yet a hint that a particular configuration will be confusing is obtained if the display confused the grapher. Shown in Figure 23 is a plot of per share earnings and dividends over a six-year period. We note (with some amusement) that 1975 is the side of a bar—the third dimension of this bar (rectangular parallelepiped?) chart has confused the artist! I suspect that 1975 is really what is labeled 1976, and the unlabeled bar at the end is probably 1977. A simple line chart with this interpretation is shown in Figure 24.

In Section 4 we illustrate six more rules for displaying data badly. These rules fall broadly under the heading of how to obscure the data. The techniques mentioned were to change the scale in mid-axis, emphasize the trivial, jiggle the baseline, order the chart by a characteristic unrelated to the data, label poorly, and include more dimensions or decimal places than are justified or needed. These methods will work separately or in combination with others to produce graphs and tables of little use. Their common effect will usually be to leave the reader uninformed about the points of interest in the data, although sometimes they will misinform us; the physicians' income plot in Figure 13 is a prime example of misinformation.

Finally, the availability of color usually means that there are additional parameters that can be misused. The U.S. Census' two-variable color map is a wonderful example of how using color in a graph can seduce us

Table 2. Optimal Selection From a Finite Sequence With Sampling Cost (revised)

N	b/c = 10		b/c = 100		b/c = 1,000	
	r*	G	r*	G	r*	G
3	2	.2	2	2.2	2	22
4	2	.3	2	2.9	2	29
5	2	.3	3	3.5	3	36
6	3	.4	3	4.2	3	43
7	3	.4	3	4.9	3	49
8	3	.5	4	5.6	4	56
9	3	.6	4	6.3	4	63
10	4	.6	4	6.9	4	70

NOTE: $g(X_s | r-1) = bR(X_s | r-1) + a$, if $S = s$, and $g(X_s | r-1) = 0$, otherwise.

into thinking that we are communicating more than we are (see Fienberg 1979; Wainer and Francolini 1980; Wainer 1981). This leads us to the last rule.

Rule 12—If It Has Been Done Well in the Past, Think of Another Way to Do It

The two-variable color map was done rather well by Mayr (1874), 100 years before the U.S. Census version. He used bars of varying width and frequency to accomplish gracefully what the U.S. Census used varying saturations to do clumsily.

A particularly enlightening experience is to look carefully through the six books of graphs that William Playfair published during the period 1786–1822. One discovers clear, accurate, and data-laden graphs containing many ideas that are useful and too rarely applied today. In the course of preparing this article, I spent many hours looking at a variety of attempts to display

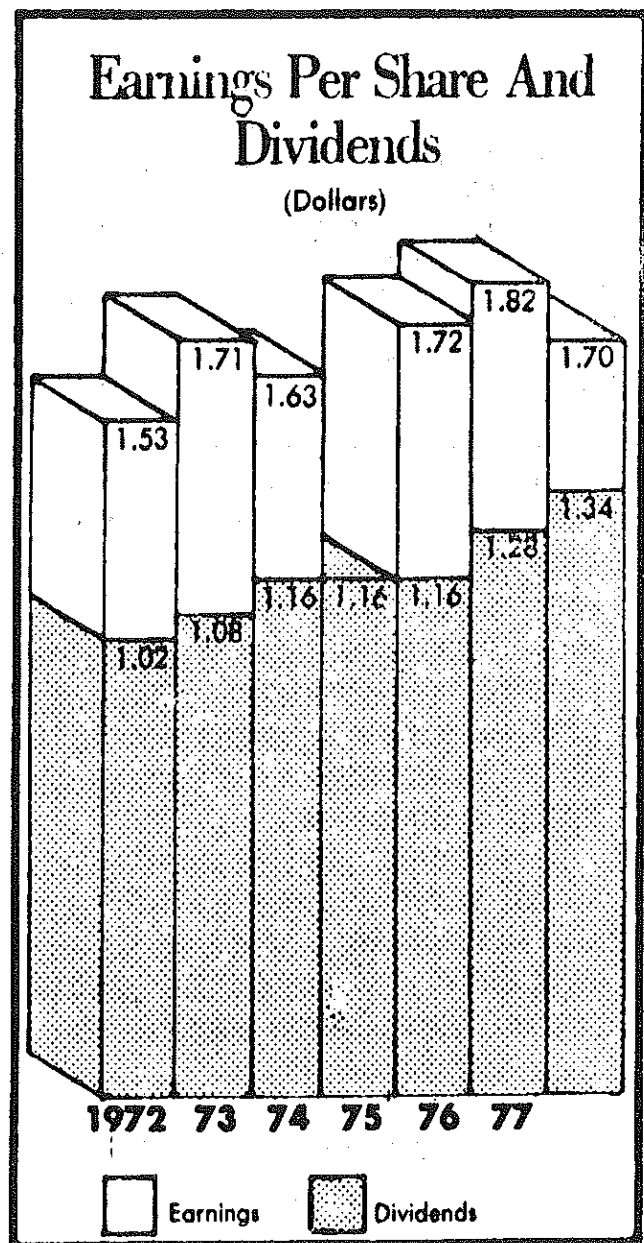


Figure 23. An extra dimension confuses even the grapher (© 1979, The Washington Post).

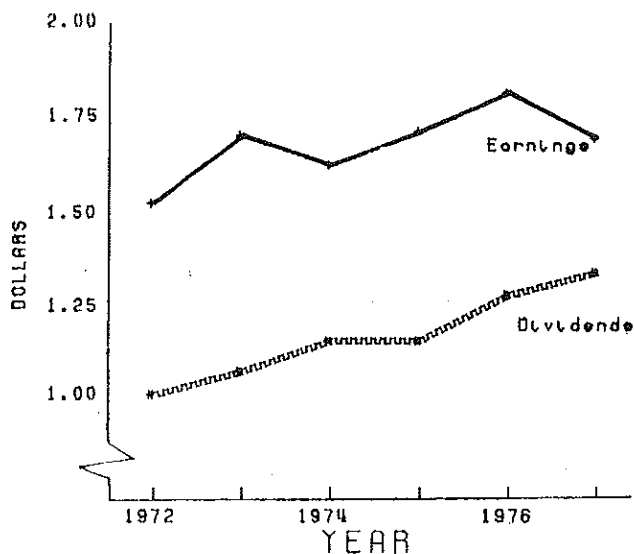


Figure 24. Data from Figure 23 redrawn simply (from Wainer 1980).

data. Some of the horrors that I have presented were the fruits of that search. In addition, jewels sometimes emerged. I saved the best for last, and will conclude with one of those jewels—my nominee for the title of "World's Champion Graph." It was produced by Minard in 1861 and portrays the devastating losses suffered by the French army during the course of Napoleon's ill-fated Russian campaign of 1812. This graph (originally in color) appears in Figure 25 and is reproduced from Tufte's book (1983, p. 40). His narrative follows.

Beginning at the left on the Polish-Russian border near the Nieman River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then sacked and deserted, with 100,000 men. The path of Napoleon's retreat from Moscow is depicted by the darker, lower band, which is linked to a temperature scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back to Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow.

It may well be the best statistical graphic ever drawn.

5. SUMMING UP

Although the tone of this presentation tended to be light and pointed in the wrong direction, the aim is serious. There are many paths that one can follow that will cause deteriorating quality of our data displays; the 12 rules that we described were only the beginning. Nevertheless, they point clearly toward an outlook that provides many hints for good display. The measures of display described are interlocking. The data density cannot be high if the graph is cluttered with chartjunk; the data-ink ratio grows with the amount of data displayed; perceptual distortion manifests itself most fre-

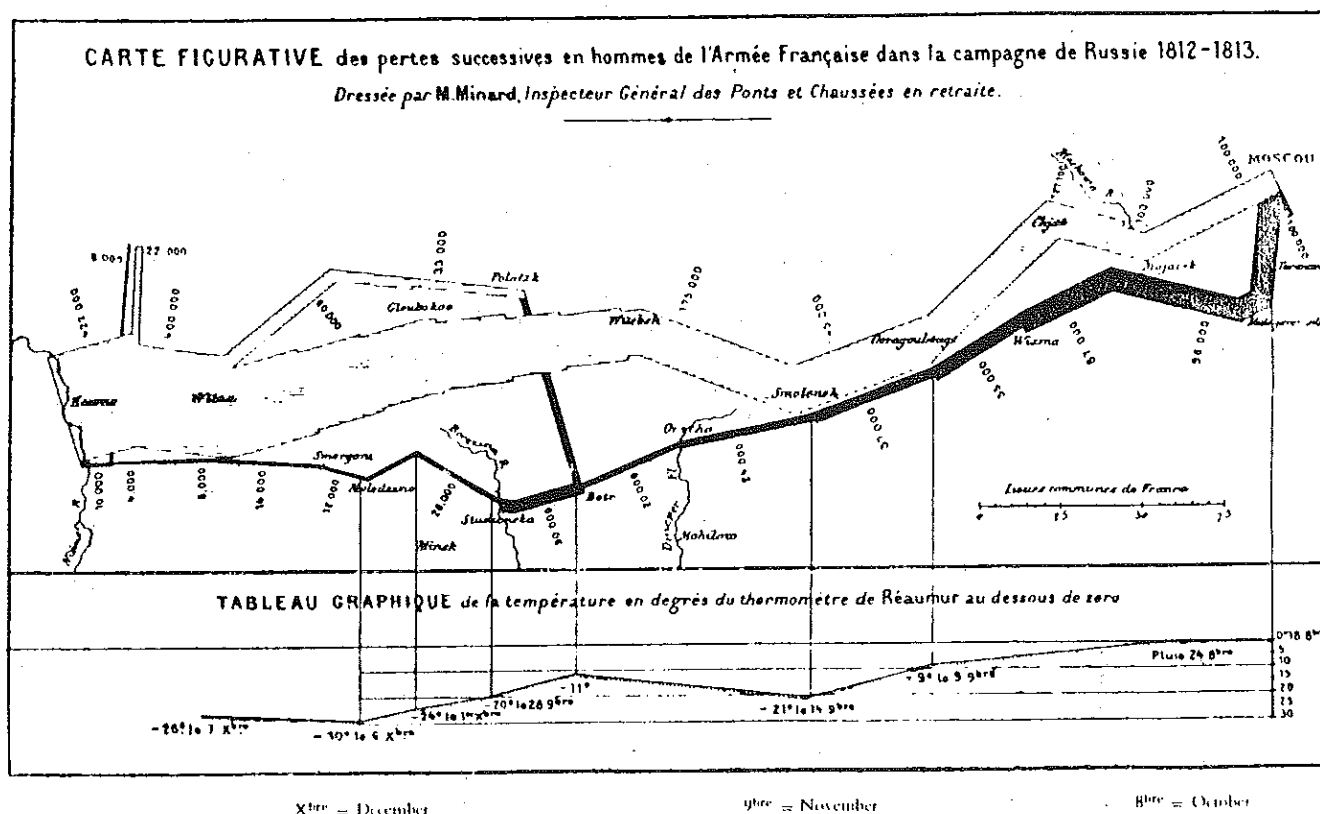


Figure 25. Minard's (1861) graph of the French Army's ill-fated foray into Russia—A candidate for the title of "World's Champion Graph" (see Tufte 1983 for a superb reproduction of this in its original color—p. 176).

quently when additional dimensions or worthless metaphors are included. Thus, the rules for good display are quite simple. Examine the data carefully enough to know what they have to say, and then let them say it with a minimum of adornment. Do this while following reasonable regularity practices in the depiction of scale, and label clearly and fully. Last, and perhaps most important, spend some time looking at the work of the masters of the craft. An hour spent with Playfair or Minard will not only benefit your graphical expertise but will also be enjoyable. Tukey (1977) offers 236 graphs and little chartjunk. The work of Francis Walker (1894) concerning statistical maps is clear and concise, and it is truly a mystery that their current counterparts do not make better use of the schema developed a century and more ago.

[Received September 1982. Revised September 1983.]

REFERENCES

- BERTIN, J. (1973), *Semiologie Graphique* (2nd ed.), The Hague: Mouton-Gautier.
- (1977), *La Graphique et le Traitement Graphique de l'Information*, France: Flammarion.
- (1981), *Graphics and the Graphical Analysis of Data*, translation, W. Berg, tech. ed., H. Wainer, Berlin: DeGruyter.
- COX, D.R. (1978), "Some Remarks on the Role in Statistics of Graphical Methods," *Applied Statistics*, 27, 4-9.
- DHARIYAL, I.D., and DUDEWICZ, E.J. (1981), "Optimal Selection From a Finite Sequence With Sampling Cost," *Journal of the American Statistical Association*, 76, 952-959.
- EHRENBERG, A.S.C. (1977) "Rudiments of Numeracy," *Journal of the Royal Statistical Society, Ser. A*, 140, 277-297.
- FIENBERG, S.E. (1979), Graphical Methods in Statistics, *The American Statistician*, 33, 165-178.
- FRIEDMAN, J.H., and RAFSKY, L.C. (1981), "Graphics for the Multivariate Two-Sample Problem," *Journal of the American Statistical Association*, 76, 277-287.
- JOINT COMMITTEE ON STANDARDS FOR GRAPHIC PRESENTATION, PRELIMINARY REPORT (1915), *Journal of the American Statistical Association*, 14, 790-797.
- MACDONALD-ROSS, M. (1977), "How Numbers Are Shown: A Review of Research on the Presentation of Quantitative Data in Texts," *Audiovisual Communications Review*, 25, 359-409.
- MAYR, G. VON (1874), "Gutachten Über die Anwendung der Graphischen und Geographischen," *Method in der Statistik*, Munich.
- MINARD, C.J. (1845-1869), *Tableaux Graphiques et Cartes Figuratives de M. Minard*, Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris.
- PLAYFAIR, W. (1786), *The Commercial and Political Atlas*, London: Corry.
- SCHMID, C.F. (1954), *Handbook of Graphic Presentation*, New York: Ronald Press.
- SCHMID, C.F., and SCHMID, S.E. (1979), *Handbook of Graphic Presentation* (2nd ed.), New York: John Wiley.
- TUFTE, E.R. (1977), "Improving Data Display," University of Chicago, Dept. of Statistics.
- (1983), *The Visual Display of Quantitative Information*, Cheshire, Conn.: Graphics Press.
- TUKEY, J.W. (1977), *Exploratory Data Analysis*, Reading, Mass: Addison-Wesley.
- WAINER, H. (1980), "Making Newspaper Graphs Fit to Print," in *Processing of Visible Language*, Vol. 2, eds. H. Bouma, P.A. Kolers, and M.E. Wrolsted, New York: Plenum, 125-142.
- , "Reply" to Meyer and Abt (1981), *The American Statistician*, 57.
- (1983), "How Are We Doing? A Review of Social Indicators III," *Journal of the American Statistical Association*, 78, 492-496.
- WAINER, H., and FRANCOLINI, C. (1980), "An Empirical Inquiry Concerning Human Understanding of Two-Variable Color Maps," *The American Statistician*, 34, 81-93.
- WAINER, H., and THISSEN, D. (1981), "Graphical Data Analysis," *Annual Review of Psychology*, 32, 191-241.
- WALKER, F.A. (1894), *Statistical Atlas of the United States Based on the Results of the Ninth Census*, Washington, D.C.: U.S. Bureau of the Census.

ACCELERATED LIFE TEST: AN OVERVIEW AND SOME RECENT ADVANCES

Gouri K. Bhattacharyya
University of Wisconsin--Madison

ABSTRACT. Statistical inferences on the durability of a product may often have to be based on an analysis of failure data generated under an overstress or accelerated life test (ALT). The effectiveness of such inferences rests heavily on the validity of model assumptions concerning the life distribution and the effect of stress acceleration. In this article, the principal methodological approaches to ALT analysis are reviewed in light of plausibility of the model, flexibility of empirical fit and usefulness in practical application. These include parametric log-linear models, semi-parametric formulations based on proportional hazards or time transformation, and a reciprocal-linear regression model in the setting of a Brownian motion process for damage growth. Some theoretical considerations and practical issues of designing an ALT experiment are also discussed.

I. INTRODUCTION. A problem frequently encountered in engineering research and development is to ascertain the durability or service life of a new product or to compare alternative designs of the same product. Usually, long life of the product and relatively much shorter time available for testing purposes impair our ability to collect failure data by conducting tests under its normal conditions of use. With accelerated life test (ALT), prototypes of the product are subjected to stress conditions that are more severe than encountered in normal use so that more failures are apt to take place in a limited time. Data of failure times under such over-stress conditions are then analyzed in the framework of a statistical model, and inferences are drawn in regard to life length or reliability of the product under its normal use condition.

Another means of reducing the test time, called censored sampling, consists of testing a larger number of units in order to observe a fewer number of failures--those that occur early. Censored life tests under normal use conditions are useful as long as failures are likely to occur within the permitted test time. When that is not the case, ALT is the only means of getting some failure data. In practice, ALT and censoring are often coupled in the same experiment toward the common goal of cost and time savings.

With technological advances leading to enhancement of product life, ALT is assuming an ever increasing role in engineering experimentation. The last two decades have seen a large growth of literature in statistical methodology for ALT analysis. The diversity of practical application has increased at the same time. A few examples are: self-lubricated bearings for high vacuum application (Meeks 1980) tested under high speed stresses, stress-rupture of Kevlar-epoxy composite (Glaser 1984) under tensile and temperature stresses, twisted nematic liquid-crystal display (Kitagawa et al 1984) under accelerated voltage stresses,

Research supported by Office of Naval Research under
Grant N00014-78-C-0722.

insulation resistance of high K multilayer ceramic capacitors (Minford 1982) under voltage and temperature stresses, and failure of power cable insulation (Lyle and Kirkland 1981) under temperature, moisture and voltage stresses. The conduct and analysis of such experiments often draw a great deal from theoretical models of chemical reaction, metal fatigue, creep rupture, wear, etc. as is relevant to the particular physical process of failure, and are aided by empirical evidence and statistical tools. The subject matter is heavily interdisciplinary, and accordingly, the relevant literature is scattered in journals of several disciplines. Our discussion will be limited to the major statistical models and methodology of ALT analysis.

To introduce the basic statistical issue of ALT we let the random variable y represent the life-length or time-to-failure of a material specimen, component or a system. The probability distribution of y depends on some identifiable environmental conditions or stresses x which are manipulated in the experiment.

Denote by x_0 the normal use-condition stress level. In an ALT experiment, a number of larger than normal stress settings x_i , $i = 1, \dots, k$ are chosen. A sample of n_i units is subjected to the constant setting x_i and either all their failure times are observed (full sample) or only some early failures are recorded (censored sample), $i = 1, \dots, k$. Thus, samples are generated from the accelerated life distributions $F(y|x_i)$, $i = 1, \dots, k$ where $F(y|x)$ denotes the cdf of y under the stress level x . Based on such data, one wishes to make inferences on some relevant characteristics of $F(y|x_0)$ such as its mean, selected percentiles, and the reliability $\bar{F}(t|x_0)$ for a mission time t where $\bar{F} = 1 - F$. Another variant, called step-stress ALT, allows the stress setting for each unit to be changed at specified intervals until failure occurs. For now we confine our attention to constant stress ALT; step-stress ALT experiments will be discussed in Section 5.

A related area of research is survival analysis in biostatistics which also deals with time (survival time, time to cure or time to onset of a disease) as the dependent variable and its dependence on such covariates as age, physiological and environmental conditions of the patient. Therefore, between ALT and survival analysis, the basic concepts, models and methods have much in common. However, considerable differences exist in regard to the conduct of the experiment, type of data, role of the covariates and the target of inference. For instance, survival analysis typically deals with a much larger set of covariates than is involved in an ALT, lesser control on the settings of the covariates, and lesser control on the process data collection which leads to more complex patterns of censoring. Also, its emphasis is toward studying the effects of some covariates after adjusting for the effects of the others -- not so much to predict $F(y|x_0)$. In fact, the concept of a normal setting for the covariates is not meaningful in survival analysis. Both of these areas can be brought under the umbrella name of

regression analysis. In essence, ALT calls for regression analysis under non-standard statistical models as well as data types, and its major goal is to make predictions beyond the range of the experimental setting. In light of the last point, it is obvious that theoretical modeling or understanding of the failure process plays a far more important role than empirical model fitting.

Inferences from ALT data require two basic ingredients of model formulation: the underlying life distribution $F(y|x)$ for a given stress x , and the functional relationship among these distributions with varying x . The

latter is sometimes called the acceleration function. The object of this paper is to give a brief survey of the various approaches to model formulation and the associated methods of statistical inference. To organize the exposition, we set out with a broad classification of the major areas of development in ALT analysis: (a) Parametric life models with log-linear acceleration function, (b) Semi-parametric approaches based on hazard rate and time-acceleration models, (c) Stochastic damage growth models, (d) Special constructs for step-stress ALT, and (e) Issues of designing an ALT experiment.

Log-linear (LL) acceleration functions in the framework of important parametric models for the underlying life distribution dominated the early developments of ALT analysis. An extensive literature has developed both in methodological advances and diverse applications. A good survey of the earlier developments is available in Chapter 9 of Mann, Schafer and Singpurwalla (MSS) (1974). The proportional hazards model, due to Cox (1972), is a semi-parametric formulation that has been found instrumental to survival analysis in biostatistics, and has led to major advances in handling arbitrarily censored data. Application of these methods to ALT is somewhat limited because the model is empirical and also the data type and object of inference are different. The semi-parametric and nonparametric approaches stem from ideas of greater generality but they typically require larger sample sizes for sensible inferences. Also, an extrapolation is less dependable when it is based on a purely empirical acceleration function. Areas of relatively recent developments include (c) and (d). For brevity, our discussion in Sections 2-5 will focus on the motivation and description of the various models and will include only an outline of the principal analytical methods. Technical details as well as treatment of special cases under each class of models will be omitted with references provided for the interested reader. Section 6 deals with designing an ALT experiment and discusses the usefulness of some optimality criteria.

2. PARAMETRIC LOG-LINEAR MODELS. A general formulation, called parametric log-linear (LL) model, consists of the following assumptions: (a) the underlying life distribution belongs to a specified parametric family involving a scale parameter θ and possibly also a shape parameter n , (b) the scale parameter depends on the stress x according to an LL-relation $\log \theta = \beta'x$ while n is

is a constant independent of x . Here x is a p-vector whose components need

not correspond to all distinct stress variables, some may be just different functions of the same variable. For instance, with temperature as the sole stress

variable x , the quadratic function $\beta_0 + \beta_1 x + \beta_2 x^2$ satisfies this formulation with $x' = (1, x, x^2)$ and $\beta' = (\beta_0, \beta_1, \beta_2)$.

The choice of a life distribution is guided by such criteria as its theoretical basis in reliability, simplicity of inference procedures and flexibility of empirical fit. Distributions derived from Poisson shocks, extreme value theory, failure rate behavior or those with good track record in fitting to life data are the natural candidates. These include the exponential, Weibull, gamma and lognormal distributions. The assumption of an LL relation to stress is not only simple and flexible but is also motivated in many practical contexts from theoretical constructs based on chemical kinetics, activation energy, principles of quantum mechanics, etc. The Arrhenius reaction rate model, Inverse power law, Eyring model, and Generalized Eyring model are some of the widely used engineering models which fit into the LL formulation. These are respectively given by

$$\begin{aligned}\theta &= \exp(A-B/x), \text{ temperature stress} \\ \theta &= (A/x)^P, \text{ voltage stress} \\ \theta &= x \exp(A-B/x), \text{ temperature stress} \\ \theta &= Ax_1 \exp(-B/x_1) \exp(Cx_2 + Dx_2/x_1), \text{ temperature and voltage stresses.}\end{aligned}\tag{2.1}$$

Statistical inferences including estimation of the model parameters and setting confidence bounds for the mean life or a specified percentile of the life distribution at use condition stress as well as model checking and goodness-of-fit are extensively treated in the literature under various distributional assumptions and specific engineering models. One general body of methodology is based on the maximum likelihood (ML) estimation, the Fisher information matrix and the associated asymptotic normal approximation. The technical details vary according to the specific models and data types, and the plethora of results are beyond the scope of this brief survey. The reader may refer to Chapter 9 of MSS (1974) for some details and also the relevant references.

In general, the maximum likelihood method in the ALT context and especially with censored data involves considerable computational complexity, and lacks a grip on the small sample properties of the estimators. Some interesting alternative procedures have been developed for the case of location-scale parameter families for the distribution of the log-life. In particular, the logarithm of Weibull and lognormal random variables have the Gumbel extreme value and normal distributions, respectively, each of which constitutes a location-scale family.

A simple estimation procedure with type II censored data, proposed by Nelson and Hahn (1972, 1973), is based on an application of the least squares method in two stages. To outline the idea we consider a p -vector \underline{x} of stress

variables with k settings x_1, \dots, x_k . At x_i , n_i units are simultaneously tested and observed till the r_i th failure occurs so for each $i = 1, \dots, k$ we have a type II right censored sample $y_{i1} < y_{i2} < \dots < y_{ir_i}$. With a minor misuse

of notation, here we take y to be the log-life so the censored sample comes from a pdf of the form $\sigma^{-1}g[(y-\lambda_i)/\sigma]$ where $\lambda_i = \beta'x_i$, g is a completely specified pdf (standard extreme-value, normal, etc.), and β and σ are the unknown parameters. For simplicity, we confine further discussion to equal sample sizes and equal censoring, that is, $n_i = n$ and $r_i = r$ for all i .

In stage 1, we ignore the regression structure and estimate the parameters (λ_i, σ) from the i th data set by the method of least squares applied to the linear model

$$y_{ij} = \lambda_i + \sigma V_{ij}, \quad j = 1, \dots, r \quad (2.2)$$

where V_{ij} , $j = 1, \dots, r$ are the first r order statistics of a sample of size n from the standardized pdf g . Their means $c_j = E(V_{ij})$ and covariances $\sigma_{jj'} = \text{cov}(V_{ij}, V_{ij'})$ are known constants, and their tables are available for some distributions. We thus have the stage-1 best linear unbiased estimators (BLUE) of the form

$$\lambda_i^* = \sum_{j=1}^r a_j y_{ij}, \quad \sigma_i^* = \sum_{j=1}^r b_j y_{ij} \quad (2.3)$$

as well as their exact covariance matrix

$$\sigma^2 \begin{pmatrix} d_1 & d_3 \\ d_3 & d_2 \end{pmatrix} \quad (2.4)$$

where d_1 , d_2 and d_3 are known constants.

In stage 2, we denote $\tilde{\lambda}^* = (\lambda_1^*, \dots, \lambda_k^*)'$, $\tilde{\sigma}^* = (\sigma_1^*, \dots, \sigma_k^*)'$, and form the linear model

$$\begin{aligned} \tilde{\lambda}^* &= \tilde{X}\beta + e_1 \\ \tilde{\sigma}^* &= \mathbf{1}\sigma + e_2 \end{aligned} \quad (2.5)$$

where $\tilde{X}' = (x_1', \dots, x_k')$, and the pair (e_1, e_2) has mean $(0, 0)$, its elements are independent across rows and have the covariance structure (2.4) across columns. Based on this linear model, the BLUE's are obtained as

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{\lambda}^*, \quad \tilde{\sigma} = \frac{1}{k} \sum_{i=1}^k \sigma_i^* \quad (2.6)$$

The mean log-life at the use condition stress x_0 as well as any percentile is of the form $\beta'x_0 + c\sigma$ which is a linear combination of β and σ .

Therefore, (2.6) leads to unbiased estimators of these quantities as well as their exact variances as opposed to only asymptotic results obtainable for the MLE's. However, to construct confidence bounds, one has to resort to large sample normal approximation of these estimators except for some isolated simple models where an exact pivotal method may be feasible, cf. McCool (1980).

Bhattacharyya and Soejoeti (1981) examine conditions on the design matrix X and the underlying log-life distribution g for the asymptotic normality of the

ML and two-stage least squares estimators, and investigate the loss of asymptotic ($k \rightarrow \infty$) efficiency incurred by the latter. In particular, for the Weibull life distribution, it is found that a fairly high efficiency is retained unless either n is too small or r is too small compared to n . Nelson (1970) discusses another two-stage estimation method where MLE is used in the first stage followed by least squares in the second but one loses the exact properties (unbiasedness, variances and covariances) in this process.

For the lognormal life model and type II right censored ALT data, Mehrotra and Bhattacharyya (1985) develop another simple and highly efficient estimation procedure using a judicious modification of the likelihood equations. Denoting

$$y_i = (y_{i1}, \dots, y_{ir_i})', \quad y' = (y_1', \dots, y_k'), \quad r = \sum_{i=1}^k r_i, \quad z_{ij} = (y_{ij} - \beta'x_{ij})/\sigma,$$

they observe that the likelihood function is a product of the two components

$$\begin{aligned} L_1 &= \sigma^{-r} \exp[-(y - X\beta)'(y - X\beta)/(2\sigma^2)] \\ L_2 &= \prod_{i=1}^k [1 - \Phi(z_{ir_i})]^{n_i - r_i} \end{aligned} \quad (2.7)$$

where X is now the $r \times p$ matrix whose rows are x_1', \dots, x_k' repeated r_1, \dots, r_k times, respectively, and Φ denotes the standard normal cdf. The factor L_1 has the form of a full sample normal regression likelihood based on the sample sizes r_1, \dots, r_k at the k design points. Complication in obtaining the MLE arises because of L_2 . A method of modified MLE is proposed by replacing $\partial \log L_2 / \partial \beta$ and $\partial \log L_2 / \partial \sigma$ by their respective expectations in the likelihood equations. It turns out that these modified likelihood equations lead to the exact solutions

$$\begin{aligned} \tilde{\beta} &= S^{-1}X'y - \tilde{\sigma}S^{-1}a \\ \tilde{\sigma}^2 &= c^{-1}y'(I - XS^{-1}X')y \end{aligned} \quad (2.8)$$

where $S = \underline{X}'\underline{X}$, and the constants a and c can be calculated by using the tables of means and variances of the standard normal order statistics. Closed form expressions, easy computing algorithm, some exact small sample properties, and little loss of asymptotic efficiency with light censoring are the principal advantages with this method. A few other modifications of the likelihood equations to obtain estimators in closed forms are discussed by Tiku (1978) and Schneider (1984) for censored normal samples.

In most applications of the parametric LL analysis, the shape parameter n is assumed to be independent of \underline{x} . Glaser (1984) employs a more general formulation with the Weibull distribution assuming that the reciprocal of the shape parameter also has a linear model in terms of \underline{x} . Iterative solution of the ML equations are discussed in the settings of grouped and censored ALT data. Shaked (1978) discusses ML estimation with the inverse power law and Arrhenius acceleration functions applied to some linear hazard rate type distributions.

3. SEMI-PARAMETRIC MODELS--PROPORTIONAL HAZARDS AND TIME ACCELERATION.

3.1 Proportional Hazards Model. The LL model discussed in the preceding section envisions a multiplicative effect of stress on the scale parameter and hence on the mean as well as the percentiles of the life distribution. Another approach to modeling the effect of stress focuses on the failure rate behavior. The failure rate at age y of a unit undergoing a constant stress \underline{x} is defined as $h(y|\underline{x}) = f(y|\underline{x})/\bar{F}(y|\underline{x})$ where f and \bar{F} are respectively the pdf and survival function of the life distribution. Let $h_0(y) = h(y|\underline{x}_0)$ denote the failure rate function under the use condition stress \underline{x}_0 . The proportional hazards (PH) model assumes that stress acts multiplicatively on the failure rate, that is $h(y|\underline{x}) = h_0(y)g(\underline{x}, \underline{\beta})$ where g is a positive function involving an unknown parameter vector $\underline{\beta}$ but is free of y . Cox (1972) proposed this idea and further assumed an exponential form of g ,

$$h(y|\underline{x}) = h_0(y)\exp(\underline{\beta}'\underline{x}) \quad (3.1)$$

arguing that this choice is "convenient, flexible and yet entirely empirical". The model is semi-parametric because one component, namely, the acceleration function is parameterized while the form of the use condition hazard $h_0(y)$ is left completely arbitrary.

The PH model has spurred extensive research in statistical methodology with applications targeted mainly to survival analysis in biostatistics. Also, handling arbitrary or randomly censored data has been a focal point of these developments. The parameter $\underline{\beta}$ is usually viewed as the primary target of inference while $h_0(y)$ is considered a nuisance function. In the context of ALT,

$h_0(y)$ or the corresponding life distribution $F(y|x_0)$ is of main interest while an assessment of the significance of the stress effects is often redundant. More importantly, the use of an empirical acceleration function with no physical back-up is prone to criticism because this function plays a dominant role in extrapolation and inferences on $h_0(y)$.

A comparison of the structures of the LL and PH models is in order here. The well known relations between the failure rate, cumulative hazard and survival functions (cf Kalbfleisch and Prentice 1980) lead to the following equivalent forms of the PH model:

$$\begin{aligned}\bar{F}(y|x) &= [\bar{F}(y|x_0)]^{\exp(\beta'x)} \\ \log[-\log\bar{F}(y|x)] &= \log[-\log\bar{F}(y|x_0)] + \beta'x.\end{aligned}\tag{3.2}$$

The second equation shows a linear model in regard to the influence of the stresses operates additively on the log (-log)-survival function. By contrast, the LL model assumes a linear form for the logarithm of the scale parameter, and is therefore physically more meaningful. It entails that $y|x$ has the same

distribution as that of $(y|x_0)[\exp(\beta'x)]$, and this relation leads to the failure rate relation

$$h(y|x) = \exp(\beta'x)h_0[y \exp(\beta'x)].\tag{3.3}$$

Obviously, the LL and PH models coincide if and only if $h_0(y) \propto y^\delta$, that is, the underlying life distribution is Weibull.

A more general class of models is formulated by Ciampi and Etezadi-Amoli (1985) by embedding both LL and PH failure rate functions in a common frame:

$$h(y|x) = \exp(\alpha'x)h_0[y \exp(\beta'x)].\tag{3.4}$$

This reduces to LL if $\alpha = \beta$ and to PH if $\beta = 0$. They study asymptotic

likelihood ratio tests for model discrimination under the further assumption that h_0 is a polynomial. It is not clear if such an over-parameterization is

necessary or meaningful in ALT analysis. The model being purely empirical, its use in ALT is questionable.

3.2 Time-Acceleration Model. The concept of a failure-time acceleration or shortening of the life-time under increased stress has prevailed in much of the historical developments of the ALT models. A simple formulation was advanced by Allen (1959) and its ramifications treated later by several authors. To introduce the basic idea, suppose $\bar{F}_0(y)$ and $\bar{G}(y)$ denote the survival functions under the

use condition stress and an accelerated stress condition, respectively. A relation between them is modeled as $\bar{G}(y) = \bar{F}_0[v(y)]$ with a "time-acceleration"

function $v(y)$. Allen (1959) calls it a strict acceleration if $v(y) > y$ for

all y (it is understood that $v(y)$ is not identically equal to y), and a restricted acceleration if $v(y) < y$ holds on a finite interval and $v(y) > y$ on an infinite interval. Note that a strict acceleration is equivalent to the use condition life being stochastically larger than that under the accelerated stress. Barlow and Scheuer (1971) considered nonparametric estimation of F_0 and G under the assumptions that both are IFRA distributions, $v(t)$ is arbitrary, and data are available from both F_0 and G .

Lacking data from F_0 , as is usually the case with ALT experiments, one must specify a structure of $v(t)$ to be able to estimate F_0 . A semi-parametric formulation, proposed by Shaked, Zimmer and Ball (SZB) (1979), assumes that the stress x acts on the survival by means of a change of the time scale,

$$v(y) = g(x, \beta)y$$

where g is a specified function of x involving an unknown parameter β , and the distribution $F_0(y)$ is arbitrary. Note that the choice $g(x, \beta) = \exp(\beta'x)$ leads to the structure of the LL model of Section 2, the sole difference being that $F_0(y)$ is left nonparametric in the present formulation.

Consider the case of a single stress variable x and a scalar parameter β . Suppose that k accelerated stress settings x_i are used, n_i units are tested at x_i and all failure times y_{ij} , $j = 1, \dots, n_i$, $i = 1, \dots, k$ are observed. The model entails that $y|x_i$ has the same distribution as $\theta_{ii'}(y|x_{i'})$ where $\theta_{ii'} = g(x_i, \beta)/g(x_{i'}, \beta)$. Based on this observation, SZB (1979) propose a simple inference procedure along the following steps:

- (i) Using the data from each pair of stress settings $(x_i, x_{i'})$, obtain a consistent estimator $\hat{\theta}_{ii'}$ of the ratio of scales

$$\text{such as } \hat{\theta}_{ii'} = \bar{y}_i / \bar{y}_{i'}, \text{ where } \bar{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}.$$

- (ii) Obtain $\hat{\beta}_{ii'}$ by solving the equation
$$\hat{\theta}_{ii'} = g(x_i, \beta) / g(x_{i'}, \beta).$$

Repeating this for all pairs get $k(k-1)/2$ estimators of β .

- (iii) Form the pooled estimator $\hat{\beta} = \sum_{1 \leq i < i' \leq k} w_{ii'} \hat{\beta}_{ii'}$ using the weights $w_{ii'}$ inversely proportional to the asymptotic variances of $\hat{\beta}_{ii'}$.

- (iv) Rescale the observed failure times to pseudo-values at the use condition stress:

$$y_{ij}^* = \frac{g(x_0, \hat{\beta})}{g(x_i, \hat{\beta})} y_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, k.$$

- (v) Act as if these pseudo-values constitute a random sample of size $N = \sum_{i=1}^k n_i$ from the distribution $F_0(y)$ in order to estimate the mean, percentiles or other features of F_0 or even the whole function $F_0(y)$.

Shaked and Singpurwalla (1982) discuss goodness-of-fit tests along these lines. The appealing features of the above procedure are its simplicity which is an attraction to the practitioners, and avoidance of the assumption of a specific parametric life distribution as is involved in the LL analysis. However, large sample sizes are needed for its validity, and that is in essence a price to be paid to forego a parametric assumption. Like the LL model it does have a parametric assumption for the acceleration function and that plays a crucial role in extrapolation. In light of this, whether one chooses a flexible parametric family for $F_0(y)$ or leaves it nonparametric is not of much practical import in model fitting and inference.

Proschan and Singpurwalla (1979, 1980) discuss a Bayesian approach which circumvents the need for choosing a specific parametric acceleration function as well as the form of the life distribution. However, they assume that prior information in regard to the average failure rates over disjoint time intervals under each accelerated condition is available, and that least squares fit of a linear relation among the posterior average failure rates can be extended to the use condition stress.

4. STOCHASTIC DAMAGE GROWTH -- AN INVERSE GAUSSIAN REGRESSION MODEL. In this section, we discuss a parametric approach based on a life distribution which derives from a stochastic model of fatigue or growth of damage in a material. In contrast with direct modeling of the time-acceleration function or the failure rate behavior discussed in the previous sections, here the rate parameter of the damage growth process is modeled in relation to the stress.

Specifically, we assume that given a constant operating environment, depletion of strength or growth of damage of a material specimen over time follows

a Brownian motion process with drift $\mu > 0$ and diffusion constant δ^2 , and that the material fails when the accumulated damage exceeds a critical level $\omega > 0$. Let $X(t)$ denote the accumulated damage during the time interval $[0, t]$. The time-to-failure is then given by $y = \inf\{t: X(t) > \omega\}$ which is the first passage time of the process across ω . The above assumptions lead to the following pdf of y :

$$f(y) = (2\pi\sigma y)^{-1/2} \exp\left[-\left(\frac{y}{\theta} - 1\right)^2 / (2\sigma y)\right], \quad 0 < y < \infty \quad (4.1)$$

where $\theta = \omega/\mu$, $\sigma = \delta^2/\omega^2$, mean = θ , and variance = $\theta^3 \sigma$. This distribution is known as a Gaussian first passage time distribution in the stochastic processes literature, and is more commonly called the inverse Gaussian distribution, $IG(\theta, \sigma)$, in the statistical literature. Its analogy with and advantages over the Birnbaum-Saunders (1969) fatigue life distribution is discussed by Bhattacharyya and Fries (BJ) (1982b).

In the context of ALT, the parameter μ , which represents the mean damage growth per unit of time, is the natural choice for constructing an acceleration function in relation to the stress \underline{x} . A simple and flexible formulation due to BF (1982a, 1986) postulates a linear regression model for μ and assumes ω and δ^2 to be constants independent of \underline{x} . The latter assumption is in the spirit of the homoscedasticity assumption in the normal theory regression analysis. Thus, the distribution of the failure time under stress \underline{x} , $y|\underline{x}$, is taken to be $IG(\theta(\underline{x}), \sigma)$ whose mean $\theta(\underline{x})$ depends on the stress \underline{x} (a p-vector) according to the reciprocal-linear model $\theta^{-1}(\underline{x}) = \underline{\beta}'\underline{x}$, and σ is independent of \underline{x} .

To discuss statistical inferences with the above model, we consider an ALT experiment with k settings of \underline{x} , and a random sample of n_i failure times y_{ij} , $j = 1, \dots, n_i$ observed at the setting \underline{x}_i , $i = 1, \dots, k$. Let N , \bar{y}_i , \bar{y} respectively denote the total sample size, the i th sample mean and the grand mean, $R = N^{-1} \sum_{ij} y_{ij}^{-1}$, the grand mean of reciprocals of the observations, $V = \sum_{ij} (y_{ij}^{-1} - \bar{y}^{-1})^2$, the total reciprocal deviation, and define the matrices

$$\begin{aligned} \underline{D} &= \text{diag}(\bar{y}_1, \dots, \bar{y}_k) \quad , \quad \underline{C} = \text{diag}(n_1, \dots, n_k) \\ \underline{X}' &= (\underline{x}_1, \dots, \underline{x}_k) \quad , \quad \underline{S} = \underline{X}' \underline{C} \underline{D} \underline{X} . \end{aligned}$$

Referring to (4.1) and the regression model $\theta_i^{-1} = \underline{x}_i' \underline{\beta}$, the likelihood function L can be written in the form

$$L \propto \sigma^{-N/2} \exp\left[-\frac{1}{2\sigma} Q(\underline{\beta})\right] \quad (4.2)$$

where

$$Q(\underline{\beta}) = (\underline{D} \underline{X} \underline{\beta} - \underline{1})' \underline{C} \underline{D}^{-1} (\underline{D} \underline{X} \underline{\beta} - \underline{1}) + V . \quad (4.3)$$

From (4.2) and (4.3), BF (1986) show that the unique roots of the likelihood equations,

$$\hat{\underline{\beta}} = \underline{S}^{-1} \sum_{ij} \underline{x}_i y_{ij}^{-1} \quad , \quad \hat{\sigma} = N^{-1} Q(\hat{\underline{\beta}}) \quad (4.4)$$

provide efficient likelihood estimators, that is, they are consistent, asymptotically normal and asymptotically equivalent to the MLE's. They further exploit some convenient features of the likelihood function to arrive at an analysis of reciprocals (ANOR) table along the ideas of the analysis of variance table in the normal theory linear model analysis. The ANOR table rests on the decomposition of the total corrected sum of reciprocals

$$V = Q_{\text{Reg}} + Q_L + Q_E$$

where the components on the righthand side measure the contributions due to regression, lack of fit, and pure error, respectively, and are given by

$$\begin{aligned} Q_{\text{Reg}} &= N(\bar{\mathbf{x}}' \hat{\boldsymbol{\beta}} - \bar{y}^{-1}) \\ Q_L &= \sum_i n_i (\bar{y}_i^{-1} - \mathbf{x}_i' \hat{\boldsymbol{\beta}}) \\ Q_E &= \sum_{ij} (y_{ij}^{-1} - \bar{y}_i^{-1}) \end{aligned}$$

Consideration of likelihood ratio tests along with a judicious intermix of exact distribution theory of IG and asymptotic theory further lead to approximate F tests for the relevant hypotheses.

Other developments in the area of IG reciprocal linear model include: construction of standardized IG residuals and their plots for a graphical model checking, construction of unbiased estimators via least squares applied to the reciprocals (BF 1982b), determination of optimal designs by minimizing a finite sample version of the asymptotic generalized variance (Fries and Bhattacharyya (FB) 1986), and analysis of factorial life test experiments (FB 1983).

The method of ALT analysis discussed in this section rests on a parametric formulation much in line with the model presented in Section 2. The IG distribution as a life model has a sound theoretical basis, and the family is flexible enough to fit most real life data just as the lognormal and Weibull families. Moreover, the reciprocal linear model as an acceleration function derives from a plausible assumption about the damage caused by stress. Taken together, the methodology of this section has several desirable features: a physical basis of the model, flexibility of empirical fit, tractability of statistical inferences and availability of model checking procedures. However, simple methods of statistical inferences with censored data are still not available for this model and further work in this direction is needed.

5. STEP-STRESS ALT. The preceding sections were concerned with the ALT studies where each unit is subjected to a constant level of stress until failure occurs or the observation is censored. Another widely used method of conducting an ALT experiment, called a step-stress ALT, allows the stress setting of a unit to be changed at discrete points of time. Stress changes may be effected at preset times or upon occurrence of a fixed number of failures along the ideas of type I and type II censoring, respectively. Applications of step-stress ALT are cited by Nelson (1980), Bora (1979) and Miller and Nelson (1983) in the contexts of failure of cable insulation under voltage stress, life testing of diodes, and dielectric breakdown of insulating fluid, respectively.

In an ordinary fixed-time step-stress experiment, a random sample of N units are simultaneously exposed to a stress setting x_1 , observed over a fixed time t_1 and the failure times of those failing in this interval are recorded.

At time t_1 , the surviving units are subjected to a different stress setting x_2 and observed till they all fail. Such an experiment is called a two-step or simple step-stress ALT. The idea extends to more than two steps in an obvious way. Moreover, the failure observations at the terminal step may be censored at a fixed time. The intent of such an experiment is to collect more failure time data in a limited time horizon without necessarily using high stresses to all the units. With an initial low stress, a unit may tend to survive too long in which case observation of its actual failure time would be lost due to censoring. That can be prevented by increasing the stress at an intermediate point thus increasing the chance of an early failure. In principle, an initial high stress can be followed by a lower one in the second step but the motivation of using this pattern is not transparent.

As with a constant stress experiment, the goal of statistical analysis of step-stress ALT data is to draw inferences on $F_0(y) = F(y|x_0)$, the life

distribution corresponding to the constant use condition stress x_0 . For this

to be possible, we must have a model that relates the step-stress life distribution to the constant stress life distribution $F_0(y)$. A sensible

formulation, called a cumulative exposure (CE) model, was proposed by Nelson (1980). It assumes that "the remaining life of specimens depends only on the current cumulative fraction failed and current stress -- regardless how the fraction accumulated. Moreover, if held at the current stress, survivors will fail according to the cdf for that stress but starting at the previously accumulated fraction failed." To formalize this idea, we let $F_i(y)$ stand for

$F(y|x_i)$, the life distribution under the constant stress x_i , and let $G(y)$

denote the life distribution under a two-step (first x_1 and then x_2) stress. The CE model entails that

$$\begin{aligned} G(y) &= F_1(y) && \text{for } y < t_1 \\ &= F_2(s_1 + y - t_1) && \text{for } t_1 < y < \infty \end{aligned} \quad (5.1)$$

where s_1 is the solution of $F_2(s_1) = F_1(t_1)$. Initially, G is the same as F_1 . At time t_1 , it switches to the function F_2 but starting with the value $F_1(t_1)$. Thus $G(y)$ is made up of segments of the constant stress life

distributions F_1 and F_2 , pieced together at the change point of stress. Note

that this formulation is different from the mixture models as well as the change point models that appear in some areas of the statistical literature.

With the general formulation (5.1), a parametric model can be constructed by taking F_1 and F_2 to be members of a common parametric family along with an LL model of relation between them. For example, use of the Weibull model

$\bar{F}(y|x) = \exp[-(y/\theta(x))^\beta]$ in conjunction with the inverse power law $\theta(x) = (A/x)^P$ and equation (5.1) leads to the step-stress life distribution

$$\begin{aligned}\bar{G}(y) &= e^{-[y(x/A)^P]^\beta}, & 0 < y < t_1 \\ &= e^{-[t_1(x_1/A)^P + (y-t_1)(x_2/A)^P]^\beta}, & t_1 < y < \infty\end{aligned}\quad (5.2)$$

where A , P and β are unknown parameters. Nelson (1980) and Miller and Nelson (1983) discuss maximum likelihood estimation under this type of parametric models where the underlying life distribution is taken to be exponential or Weibull, and the acceleration function either Arrhenius or the inverse power law. They also illustrate application to data of some step-stress ALT experiments.

A physical basis of the CE model in step-stress ALT is not as transparent as its mathematical formulation. Earlier, in a similar context, DeGroot and Goel (1979) advanced a time-acceleration model which is physically more meaningful. They assume that the effect of switching the stress from x_1 to x_2 is to

multiply the remaining life of the unit by some unknown factor α , a function of x_1 and x_2 ($\alpha < 1$ if x_2 is more severe than x_1). Letting y_1 denote the life-length under the constant stress x_1 and y^* that under the step-stress pattern (switching from x_1 to x_2 at time t_1), they formulate the relation

$$\begin{aligned}y^* &= y_1 & \text{if } y_1 < t_1 \\ &= t_1 + \alpha(y_1 - t_1) & \text{if } t_1 < y_1\end{aligned}\quad (5.3)$$

and call y^* a tampered random variable. It can be seen that (5.3) becomes a special case of (5.1) if F_1 and F_2 differ only by a scale parameter. In this

sense, (5.1) accommodates a more general formulation by allowing other parameters of the life distribution to change with stress, although such a generalization obscures the physical meaning of the model and in none of the applications it has been used as yet. DeGroot and Goel (1979) only consider the setting of a "partially accelerated life test" viewing x_1 as the use condition stress and

x_2 the single accelerated stress so a specification of the acceleration function, relating α to x , is not necessary. On the other hand, they allow t_1 to be

different for different units. Considering the underlying life distribution to be exponential, they study the issue of optimal design in the framework of Bayesian decision theory along with the specification of some cost function. Goel (1975) discusses the asymptotic properties of MLE in the above setting.

Curiously, with the assumption of an exponential distribution but without any reference to ALT, the above model also appears in the literature under a different name -- a change point hazard model. The formulation, which is in terms of the failure rate function is

$$\begin{aligned} h(y) &= \lambda_1 \text{ if } y < t_1 \\ &= \lambda_2 \text{ if } y > t_1, \end{aligned} \quad (5.4)$$

and it leads to the life distribution

$$\begin{aligned} g(y) &= \lambda_1 e^{-\lambda_1 y} \text{ if } y < t_1 \\ &= \lambda_2 e^{-\lambda_1 t_1 - \lambda_2 (y - t_1)} \text{ if } t_1 < y < \infty. \end{aligned} \quad (5.5)$$

Except for a change of notation, it is identical with the model (5.3) of a tampered exponential random variable. However, in the context of a change point hazard, the time point of change t_1 is regarded as an unknown parameter in

addition to the failure rates λ_1 and λ_2 . Here, the standard asymptotic theory of MLE does not apply. In fact, one faces the problem of non-existence of the MLE. Nguyen et al (1984) and Matthews and Farewell (1982) discuss parameter estimation, and testing the hypothesis of no change, and also provide references to earlier works in this area.

6. DESIGNING AN ALT. A carefully planned life test experiment is at the heart of success in gathering informative data, coping with the constraints of cost and time, and arriving at effective inferences as well as identifying directions of further investigation. Among many issues involved in planning an ALT experiment, some are to be resolved from an understanding of the physics of failure. These include choice of the stress variable(s), choice of the acceleration function consistent with a physical model of the failure process, and decision regarding the range of stress acceleration which would be feasible and dependable for the purposes of extrapolation. Moreover, accepted engineering practice in a given context should guide to the choice between a constant stress ALT and a step-stress ALT experiment.

Consider the most common type of ALT where a single stress x is accelerated, and denote by x_L and x_H the intended lowest and highest settings

of x . As before, we denote the use condition stress by x_0 so $x_0 < x_L < x_H$.

With a constant stress ALT, one needs to determine the number k of stress settings to be used, their locations in the interval $[x_L, x_H]$, the allocation of

a given total number N of units to the various stress settings, the period of observation and the scheme of censoring. Unlike the situation of normal theory regression analysis or least squares fitting of multiple regression with complete data, a statistical treatment of optimal ALT plans is made complicated by the fact that the important parametric life distribution models do not lead to exact results for the sampling distribution of the relevant estimators or manageable expressions for their variances especially in the case of type I censored data. Faced with this pervasive difficulty, one reasonable approach to address the issue

of optimal test plans is based on large sample theory of ML estimators. Nelson and Kielpinski (1975, 1976) and Nelson and Meeker (1978) discuss several test plans along this line. Their main developments are outlined below.

The specifications involved in their development of optimal test plans include: a parametric life distribution such that the log-life conforms to a location-scale family, an engineering acceleration function that conforms to the log-linear model (such as the Arrhenius or inverse power law), a total sample size N and a common censoring time T (determined from cost and schedule constraints), and the highest stress setting x_H (to be set as high as possible subject to validity of the model). The object of inference is to estimate $\mu(x_0)$, the median log-life or more generally, $\tau_p(x_0)$, the 100p percentile of the life distribution at the use condition stress x_0 . Two kinds of test plans, the best standard plans and the optimal two-point plans are discussed in this setting.

A standard plan, so called because of its popularity among practitioners, is one that uses k equispaced stresses in a suitable transformed scale, and equal number of test units at each stress. Given k , the best standard plan seeks to determine the x_L that minimizes the asymptotic variance of $\hat{\mu}(x_0)$, the MLE of $\mu(x_0)$. An optimal two-point plan uses $k = 2$ and finds the x_L and the proportion of units π_L tested at x_L so as to minimize the asymptotic variance of $\hat{\mu}(x_0)$. To arrive at these plans for the lognormal life model, Nelson and Kielpinski (1976) start with the asymptotic theory of MLE, compute the Fisher information matrix, and use the delta method to deduce an expression for the asymptotic variance of $\hat{\mu}(x_0)$. Minimization of this function is done numerically on a computer with various input values of the model parameters and other quantities that are fixed in advance, and thereby charts are prepared for guidance to the practitioner. Nelson and Meeker (1978) discuss such plans for the case of Weibull distribution along with the inverse power law acceleration. It is found that for the case of two-point designs, the optimal plan typically allocates more units to the low stress and requires a slightly lower x_L than the best standard plan. Similar issues are also discussed by Meeker and Hahn (1977) in the context of success-failure data and a logistic regression model.

It is to be noted that a determination of these optimal plans depends on the unknown model parameters which appear in the expression for the asymptotic variance of MLE. Therefore, one must have an informed guess of the parameter values either from experience with similar experiments or by conducting a preliminary ALT experiment. Also, a drawback of the two-point plans is that their optimality rests on the correct choice of the model and, at the same time, they provide little scope of checking lack-of-fit or violation of the model assumptions. To remedy this drawback without departing too much from optimality, best compromise plans are suggested. A compromise plan uses a third design point

x_M intermediate between x_L and x_H , a small proportion of units tested at x_M , and retains the same relative allocation to x_L and x_H as with the optimal plan.

Meeker (1984) reports an extensive simulation study for the purpose of comparing the above plans along with a few others determined from such requirements as equal expected number of failures rather than equal sample sizes at the design points and minimization of the variance of some other parameter estimates. The principal criteria used in this comparison include: quality of estimation under the chosen model (precision), ability to detect a departure from the assumed linear model (goodness-of-fit), sensitivity to misspecified parameter values (robustness) and ability to generate adequate failure data at the design points (feasibility). It is found that the ALT plans that are theoretically optimal have serious drawbacks in regard to the other criteria. The compromise plans are sub-optimal but are more robust and are also capable of detecting departures from the assumed model.

The above discussion summarizes the recent developments on ALT designs for the case of type I censoring scheme and parametric log linear analysis. Earlier works were confined largely to uncensored data under the exponential model with some specific acceleration function (Chernoff 1962) or the standard least squares fitting of multiple regression (Herzberg and Cox 1972). For the Weibull distribution with a polynomial function for the log-scale parameter, Mann (1972) discusses optimal test plans for estimating $\tau_p(x_0)$ by means of a linear function

of order statistics rather than the MLE. Fries and Bhattacharyya (1985) study optimal ALT designs under the inverse Gaussian distribution along with a reciprocal-polynomial regression model.

Derringer (1982) points out that in order to observe failures with a single accelerated stress, one often requires the settings so large that validity of the assumed model becomes questionable. To remedy the danger of a long-range extrapolation, he suggests the use of multiple stress acceleration so each stress factor could be employed at relatively low levels and yet together they would accomplish the purpose of a single large stress. This is also logical from a practical viewpoint because most materials or systems are affected by several stresses in their normal operation. However, with multiple stress acceleration one needs to be concerned about possible interaction of the stresses. At the same time, theoretical modeling of the acceleration function is typically more difficult when several stresses are to be accelerated simultaneously. In essence, the choice will really be between using a less reliable model for a short-range extrapolation and a more reliable model for a long-range extrapolation. For an effective resolution of such issues there ought to be sufficient interaction of the statistician with materials scientists and engineers who are knowledgeable about the mechanics of the failure process.

REFERENCES

- Allen, W. R. (1959), Inference from tests with continuously increasing stress, Journal of the Operations Research Society of America, 303-312.
- Barlow, R. E. and Scheuer, E. M. (1971), Estimation from accelerated life tests, Technometrics, 13, 145-159.
- Bhattacharyya, G. K. and Fries, A. (1982a), Inverse Gaussian regression and accelerated life tests, in Proceedings of the Special Topics Meeting on Survival Analysis, Institute of Mathematical Statistics, 101-117.
- _____, (1982b), Fatigue failure models--the Birnbaum Saunders versus the inverse Gaussian, IEEE Transactions on Reliability, R-31, 439-441.
- _____, (1986), On the inverse Gaussian multiple regression and model checking procedures,
- Bhattacharyya, G. K. and Soejoeti, Z. (1981), On the performance of least squares estimators in type-II censored accelerated life tests, IAPQR Transactions -- Jour. Ind. Assoc. for Productivity, Quality and Reliability, 6, No. 1, 39-55.
- Birnbaum, Z. W. and Saunders, S. C. (1969), A new family of life distributions, Journal of Applied Probability, 6, 319-327.
- Bora, J. S. (1979), Step-stress accelerated life testing of diodes, Microelectron Reliability, 19, 279-280.
- Ciampi, A. and Etezadi-Amoli, J. (1985), A general model for testing the proportional hazards and the accelerated failure time hypotheses in the analysis of censored survival data with covariates, Communications in Statistics -- Theory & Methods, 14(3), 651-667.
- Chernoff, H. (1962), Optimal accelerated life designs for estimation, Technometrics, 4, 381-408.
- Cox, D. R. (1972), Regression models and life tables (with Discussion), Journal of the Royal Statistical Society, B, 34, 187-220.
- DeGroot, M. H. and Goel, P. K. (1979), Bayesian estimation and optimal designs in partially accelerated life testing, Naval Research Logistics Quarterly, 26, 223-235.
- Derringer, G. (1982), Considerations in single and multiple stress accelerated life testing, Journal of Quality Technology, 14, 130-134.
- Fries, A. and Bhattacharyya, G. K. (1983), Analysis of two-factor experiments under an inverse Gaussian model, Journal of American Statistical Association, 78, 820-826.

- (1986), Optimal design for an inverse Gaussian regression model, Submitted to Statistics and Probability Letters.
- Glaser, R. E. (1984), Estimation for a Weibull accelerated life testing model, Naval Research Logistics Quarterly, 31, 559-570.
- Goel, P. K. (1975), Consistency and asymptotic normality of maximum likelihood estimators, Scandinavian Actuarial Journal, 2, 109-118.
- Herzberg, A. M. and Cox, D. R. (1972), Some optimal designs for interpolation and extrapolation, Biometrika, 59, 551-561.
- Kalbfleisch, J. D. and Prentice, R. L. (1980), The Statistical Analysis of Failure Time Data, John Wiley & Sons, New York.
- Kitagawa, K., Toriama, K. and Kanuma, Y. (1984), Reliability of liquid crystal display, IEEE Transactions on Reliability, R-33, No. 3, 213-218.
- Lyle, R. and Kirkland, J. W. (1981), Accelerated life tests for evaluating power cable insulation, IEEE Transactions Power Appar. Syst., PAS-100, No. 8, 3764-3774.
- Mann, N. R. (1972), Design of over-stress life-test experiments when failure times have the two-parameter Weibull distribution, Technometrics, 14, 437-451.
- Mann, N. R., Schafer, R. E. and Singpurwalla, N. D. (1974), Methods for Statistical Analysis of Reliability and Life Data, John Wiley & Sons, New York.
- Matthews, D. E. and Farewell, V. T. (1982), On testing for a constant hazard against a change-point alternative, Biometrics, 38, 463-468.
- McCool, J. I. (1980), Confidence limits for Weibull regression with censored data, IEEE Transactions on Reliability, R-29, No. 2, 145-150.
- Meeker, W. Q. (1984), A comparison of accelerated life test plans for Weibull and lognormal distributions and type I censoring, Technometrics, 26, 157-171.
- Meeker, W. Q. and Hahn, G. J. (1977), Asymptotically optimum over-stress tests to estimate the survival probability at a condition with a low expected failure probability, Technometrics, 19, 381-399.
- Meeks, Crawford R. (1981), Theory and practice of self-lubricated oscillatory bearings for high-vacuum applications, Part II - Accelerated life tests and analysis of bearings, Journal of the American Society of Lubrication Engineers, 37, 657-667.
- Mehrotra, K. G. and Bhattacharyya, G. K. (1985), Estimation of linear regression parameters under type-II censoring, Tech. Report No. 771, Dept. of Statistics, Univ. of Wisconsin-Madison.

- Miller, R. and Nelson, W. (1983), Optimum simple step-stress plans for accelerated life testing, IEEE Transactions on Reliability, R-32, 59-65.
- Minford, W. J. (1982), Accelerated life testing and reliability of high K multilayer ceramic capacitors, IEEE Transactions Comp. Hybrids Manuf. Tech., CHMT-5, No. 3, 297-300.
- Nelson, W. (1970), Statistical methods for accelerated life test data -- the inverse power law model, General Electric Corporate Research and Development TIS Report 71-C-120.
- ____ (1973), Analysis of residuals from censored data, Technometrics, 15, 697-715.
- ____ (1980), Accelerated life testing -- step-stress models and data analysis, IEEE Transactions on Reliability, R-29, 103-108.
- Nelson, W. and Hahn, G. J. (1972), Linear estimation of a regression relationship from censored data -- Part 1. Simple methods and their application, Technometrics, 14, 247-269.
- ____ (1973), Linear estimation of a regression relationship from censored data -- Part 2. Best linear unbiased estimation and theory, Technometrics, 15, 133-150.
- Nelson, W. and Kielpinski (1975), Optimum accelerated life tests for normal and lognormal life distributions, IEEE Transactions on Reliability, R-24, 310-320.
- ____ (1976), Theory for optimum accelerated life tests for normal and lognormal distributions, Technometrics, 18, 105-114.
- Nelson, W. and Meeker, W. Q. (1978), Theory for optimum accelerated censored life tests for Weibull and extreme value distributions, Technometrics, 20, 171-177.
- Nguyen, H. T., Rogers, G. S. and Walker, E. A. (1984), Estimation in change-point hazard rate models, Biometrika, 71, 299-304.
- Proschan, F. and Singpurwalla, N. D. (1979), Accelerated life testing -- a pragmatic Bayesian approach, Optimization in Statistics, Ed. by J. Rustagi, Academic Press, New York.
- ____ (1980), A new approach to inference from accelerated life tests, IEEE Transactions on Reliability, R-29, 98-102.
- Schneider, H. (1984), Simple and highly efficient estimators for censored normal samples, Biometrika, 71, 412-414.
- Shaked, M. (1978), Accelerated life testing for a class of linear hazard rate type distributions, Technometrics, 20, 457-466.

Shaked, M. and Singpurwalla, N. D. (1982), Nonparametric estimation and goodness-of-fit testing of hypotheses for distributions in accelerated life testing, IEEE Transactions on Reliability, R-31, 69-74.

Shaked, M., Zimmer, W. J. and Ball, C. A. (1979), A nonparametric approach to accelerated life testing, Journal of American Statistical Association, 74, 694-699.

Tiku, M. L. (1978), Linear regression model with censored observations, Communications in Statistics -- Theory and Methods, 7, 1219-1232.

Thirty-First Conference on the Design of Experiments
in Army Research, Development and Testing

23-25 October 1985

Roster of Attendees

Sung K. Ahn
UW-Madison, Statistics Dept.
1210 W. Dayton St.
Madison, WI 53706

Peter Arzberger
UW-Madison, Statistics Dept. (Same as 1)

William E. Baker
US Army Ballistic Research Laboratory
ATTN: AMXBR-SECAD
Aberdeen Proving Ground, MD 21005-5066

Carl B. Bates
US Army Concepts Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20814

Bobby Bennet
US Army Material Systems Analysis Agency
ATTN: AMXSY-RV
Aberdeen Proving Ground, MD 21005

Gouri K. Bhattacharyya
UW-Madison, Statistics Dept. (Same as 1)

Barney Bissinger
Penn State University/Hershey Foods
Mathematical Sciences
Middletown, PA 17057

Barry A. Bodt
US Army Ballistic Research Laboratory
ATTN: AMXBR-SECAD
Aberdeen Proving Ground, MD 21005-5066

Jane M. Booker
Los Alamos National Laboratory
P. O. Box 1663 - Mail Stop F600
Los Alamos, New Mexico 87545

Carl de Boer, UW-Madison, MRC & Mathematics Dept.
610 Walnut Street
Madison, WI 53705

George E. P. Box, UW-Madison, MRC & Statistics Dept.
(Same as 1)

Melvin Brown
US Army Research Office
ATTN: SLCRO-MA
Research Triangle Park, NC 27709

Richard M. Brugger
US Army Armament, Munitions and Chemical Command,
ATTN: AMSMC-QAL-N
Rock Island, IL 61299-6000

Kyung-Joon Cha, UW-Madison, Statistics Dept.
(Same as 1)

Jagdish Chandra
Director, Mathematics Division
US Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709

Zehua Chen, UW-Madison, Statistics Dept. (Same as 1)

Alan Currit
Systems Product Division, IBM
Rochester, NY

Gordon Douglas
US Army Electronic Proving Ground
ATTN: STEEP-MT-E
Fort Huachuca, AZ 85613-7110

N. R. Draper, UW-Madison, MRC & Statistics Dept.
(Same as 1)

Richard H. Duncan
US Army White Sands Missile Range
ATTN: STEWS-SC
White Sands Missile Range, NM 88002-5003

Eugene F. Dutoit
US Army Infantry School
ATTN: ATSH-CD-CS-OR
Fort Benning, GA 31905

Oskar M. Essenwanger
Commander, US Army Missile Command
Research, Development & Engineering Center
Redstone Arsenal, AL 35898-5248

Edward W. Frees, UW-Madison, School of Business
1155 Observatory Drive
Madison, WI 53706

Chong Gu, UW-Madison, Statistics Dept. (Same as 1)

James M. Hardin
AD/KRBAS, Freeman Math. Lab.
Eglin AFB, FL 32542

Bernard Harris, UW-Madison, MRC & Statistics Dept.
(Same as 1)

James B. Hofmann
AMSAA
Aberdeen Proving Ground, MD 21005

William R. Holt
US Army Aeromedical Research Laboratory
ATTN: SGRD-UAR (Mr. Holt)
P. O. Box 577
Fort Rucker, AL 36362-5000

Mei-Moun Huang, UW-Madison, Statistics Dept.
(Same as 1)

Charles J. Hunter
Directorate Mathematics & Statistics (DMS)
Department of National Defence
Ottawa, Canada K1A 0K2

Richard Johnson, UW-Madison, MRC & Statistics Dept.
(Same as 1)

Ronald L. Johnson, US Army Belvoir R&D Center

Rickey A. Kolb
United States Military Academy
West Point, NY 10996

John Latchaw
USAMETA
Rock Island, IL 61299-7040

Robert L. Launer
US Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

John Lawrence
Department of Defense
Fort George Meade, MD 20755

Jae-June Lee, UW-Madison, Statistics Dept.
(Same as 1)

Siegfried H. Lehnigk
Commander
US Army Missile Command
ATTN: AMSMI-RD-RE-OP/S. H. Lehnigk
Redstone Arsenal, AL 35898-5248

Wei-Yin Loh, UW-Madison, Statistics Dept. (Same as 1)

Feng-shi MA, University of Tianjin, China
& UW-Madison, Statistics Dept. (Same as 1)

Bard Mansager
USACDEC, ATTN: ATEC-TP
Fort Ord, CA 93941-7000

Mary Meyer
MS F611
Los Alamos National Laboratory
Los Alamos, NM 87545

Robert E. Miller
Walter Reed Army Institute of Research
Washington, DC 20012

Donald Neal
Army Materials & Mechanics Research Center

Christopher Neubert, US Army Engineer School

John W. Ogren
USACDEC, ATTN: ATEC-TP
Fort Ord, CA 93941-7000

Emanuel Parzen
Statistics Department
Texas A&M University
College Station, TX 77843

Daryl Pregibon
Bell Laboratories
Murray Hill, NJ 07974

Marion R. Reynolds, Jr.
Department of Statistics
Virginia Tech
Blacksburg, VA 24061

Paul A. Roediger
US Army Armament, Munitions and Chemical Command,
Dover, NJ 07801-5001

Carl T. Russell
USA Operational Test & Evaluation Agency
ATTN: CSTE-SP-M (Dr. Russell)
5600 Columbia Pike
Falls Church, VA 22041

Jerome Sacks
Department of Statistics
University of Illinois
1409 W. Green
Urbana, IL 61801

Stanley Sclove
Information & Decision Sciences Department
College of Business Administration
University of Illinois at Chicago
Box 4348
Chicago, IL 60680-4348

N. D. Singpurwalla
Department of Operations Research
The George Washington University
Washington, DC 20052

Andrew P. Soms
Department of Mathematics
University of Wisconsin-Milwaukee
Milwaukee, WI 53201

Douglas B. Tang
Division of Biometrics
Walter Reed Army Institute of Research
Washington, DC 20012

Malcolm S. Taylor
US Army Ballistic Research Laboratory
ATTN: AMXBR-SECAD
Aberdeen Proving Ground, MD 21005-5066

Deloris Testerman
US Army Combat Systems Test Activity
Performance Analysis Division
STECs-DA-PS, Building 400
Aberdeen Proving Ground, MD 21005

Jerry Thomas
US Army Ballistic Research Laboratory
ATTN: AMXBR-SECAD
Aberdeen Proving Ground, MD 21005-5066

Paul H. Thrasher
STEWs-PL-Q
WSMR, New Mexico 88002

Henry B. Tingey
Department of Mathematical Sciences
University of Delaware
Newark, DE 19716

Gary Travers
Department of Defense
Fort George Meade, MD 20755

Mark Vangel
Army Materials & Mechanics Research Center

Grace Wahba, UW-Madison, Statistics Dept. (Same as 1)

Howard Wainer, Educational Testing Service

Clarence P. Walters
Night Vision & Electro-Optics Lab.
Fort Belvoir, Virginia 22060-5677

William D. West
Scientific Advisor
USACDEC
Fort Ord, CA 93941-70000

Kuo-Tsung Wu, UW-Madison, Statistics Dept.
(Same as 1)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report 86-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROCEEDINGS OF THE THIRTY-FIRST CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report
7. AUTHOR(s)		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS		8. CONTRACT OR GRANT NUMBER(s)
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on Behalf of the Chief of Research, Development and Acquisition		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) US Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211		12. REPORT DATE June 1986
		13. NUMBER OF PAGES 280
		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. The findings in this report are not to be construed as Official Department of the Army position, unless so desig- nated by other authorized documents.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES This is a technical report from the Thirty-First Conference on the Design of Experiments in Army Research, Development and Testing. It contains most of the papers presented at that meeting. These articles treat various Army statistical and design problems.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) simulation models nonparametric two-sample tests cluster analysis and segmentation factor analysis small composite designs small sample quantal response testing human factors affecting subjective judgments reliability assessments hypothesis testing contingency table data probability density functions plotting mathematical functions effectiveness of camouflage colors confidence bounds the Lindstrom-Madden method how to display data badly accelerated life tests		